Benchmarking reservoir operation schemes for large-scale hydrological models

Jesús Casado-Rodríguez[1,a], Juliana Disperati[2], Stefania Grimaldi[1], and Peter Salamon[1]

[1]European Commission – Joint Research Centre, Ispra, Italy
[2]Fincons Group, Vimercate, Italy
[a]now at: Kajo s.r.o., Banská Bystrica, Slovakia

Abstract. There are approximately 62,000 large dams worldwide that significantly alter the hydrological regimes of most major rivers. Despite their importance, reservoirs remain poorly represented in Large-Scale Hydrological Models (LSHMs) due to the complexity of human-driven operations and a widespread lack of observational records. Consequently, reservoir routines in LSHMs must balance structural simplicity with limited data requirements. In this study, we utilize the ResOpsUS dataset to benchmark four reservoir routines of increasing complexity: LISFLOOD, CaMa-Flood, mHM, and STARFIT. We evaluate these routines across 164 reservoirs in the United States and test which target variables are most informative for parameter estimation. Our results indicate that the mHM routine consistently achieves the highest performance; however, its dependence on site-specific demand data limits its applicability at the global scale. In contrast, the CaMa-Flood routine provides a robust compromise, significantly outperforming the linear logic of LISFLOOD while maintaining parsimonious data requirements. Crucially, we find that calibrating to reservoir storage is more informative than calibrating to outflow, as it effectively captures the dynamics of both state variables. This finding paves the way for the use of satellite-derived storage products in the calibration of LSHMs. The findings of this study have been implemented in the upcoming versions of the European and Global Flood Awareness Systems (EFAS v6 and GloFAS v5).

1 Introduction

The increasing global population and expanding economic activities have amplified the pressure on freshwater resources, making human interventions—such as the construction and operation of dams and reservoirs—fundamental elements of the terrestrial water cycle (Biemans et al., 2011; Busker et al., 2019; Sadki et al., 2023; Moreno-Rodenas et al., 2025). These structures are pivotal in fulfilling diverse societal needs, including domestic and industrial water supply, irrigation, hydropower production, and flood control (Mekonnen and Hoekstra, 2012; Abeshu et al., 2023; Sadki et al., 2023; Lehner et al., 2024; Shrestha et al., 2024; Moreno-Rodenas et al., 2025). The global proliferation of dams has been unprecedented over the last century (Lehner et al., 2024; Moreno-Rodenas et al., 2025). Globally, there are approximately 62,000 large dams (heights > 15 m), cataloged by the ICOLD (International Commission on Large Dams) (Moreno-Rodenas et al., 2025), with an estimated cumulative storage capacity ranging between 7,420 km^3 (GDW v1.0) (Lehner et al., 2024) and 8,300 km^3 (Sadki et al., 2023). This volume

represents roughly 20% of the average annual river discharge to the oceans and is four times the amount of water stored in river channels worldwide (Sadki et al., 2023).

The immense scale of this intervention profoundly impacts natural hydrology by modifying the magnitude, timing, and duration of streamflows (Shin et al., 2019; Salwey et al., 2024), resulting in the fragmentation of over 60% of the world's largest rivers (Sadki et al., 2023). In the contiguous United States (CONUS), dams regulate all major rivers, providing a total storage capacity equivalent to roughly 75% of the mean annual CONUS runoff (Steyaert et al., 2022). While beneficial for human use, these operations cause significant environmental trade-offs, including the homogenization of river dynamics (Scherer and Pfister, 2016; Lehner et al., 2024), alteration of sediment and nutrient transport (Shin et al., 2019; Yassin et al., 2019), and increased open-water evaporation, which can exacerbate water scarcity (Scherer and Pfister, 2016; Yassin et al., 2019; Liu et al., 2026).

Despite their importance, the accurate representation of complex reservoir operations remains a challenge in large-scale hydrological models (LSHMs) and land-surface models (Turner et al., 2021; Sadki et al., 2023; Salwey et al., 2024). Reservoir operations are intrinsically non-linear and governed by complex, human-driven decisions related to water demand, flood risk, and environmental constraints (Coerver et al., 2018; Hanazaki et al., 2022; Steyaert et al., 2022; Salwey et al., 2024; Shrestha et al., 2024). Consequently, models lacking adequate reservoir schemes often exhibit degraded predictive performance, especially in highly regulated basins (Turner et al., 2021; Salwey et al., 2024; Steyaert et al., 2025). Historically, global models have relied on generic operational rules utilizing static characteristics (e.g., capacity, primary purpose) from datasets like the Global Reservoir and Dam (GRanD) database (Lehner et al., 2011). However, these generic approaches often fail to capture the local operating behaviors necessary for simulating realistic daily releases and storage dynamics (Turner et al., 2021). Recent research has therefore pivoted toward exploiting newly available observational data and integrating data-driven techniques, such as machine learning or statistically inferred policies, to better reflect historical decision-making (Coerver et al., 2018; Turner et al., 2021; Steyaert et al., 2025).

This research aims to improve the reservoir module within the OS LISFLOOD hydrological model (Roo et al., 2000; van der Knijff et al., 2010; Burek et al., 2013; JRC, 2026c), a core component of the Global and European Flood Awareness Systems of the Copernicus Emergency Management Services (GloFAS and EFAS) (JRC, 2026a, b). The current LISFLOOD reservoir module employs an operation scheme based on three fixed storage limits (conservation, normal, and flood pools) without explicit classification of reservoir purpose (Zajac et al., 2017). Our objective is to identify a reservoir routine suitable for global application—one that balances portability with efficacy in data-scarce environments.

Specifically, this study presents a systematic benchmark comparing the performance of four distinct reservoir schemes within large-scale hydrological models, all of which are calibrated using in-situ records. We selected schemes from the literature that represent an increasing level of complexity and varying modeling philosophies: the current LISFLOOD storage-based parameterization (Zajac et al., 2017), its evolution as implemented in CaMa-Flood (Hanazaki et al., 2022), a demand-driven approach from mHM (Shrestha et al., 2024), and a seasonality-driven statistical approach, STARFIT (Turner et al., 2021). While not exhaustive, this selection encompasses the primary modeling paradigms currently used in the field. Other models, such as DZTR (Yassin et al., 2019) or SBTS (Shen et al., 2025), could be evaluated in future work. By assessing the capability of these

schemes to simulate observed flow regulation and storage dynamics, this research provides critical guidance for the development of human–water interaction modules in the next generation of operational flood systems and LSHMs.

A novel aspect of this study is the exploration of a decoupled calibration strategy. Rather than calibrating reservoir parameters simultaneously with the rest of the hydrological model against downstream river discharge—a process often plagued by equifinality (Beven and Freer, 2001)—we define reservoir parameters in advance using specific reservoir records. These pre-determined parameters are then integrated into the hydrological model for the subsequent calibration of the remaining parameters. This approach ensures that reservoir parameters are sensitive to actual reservoir dynamics rather than being masked by gauge observations far downstream.

However, the practical implementation of this strategy faces a significant data bottleneck. While in situ records of reservoir outflow provide the most direct target for calibration, such data are frequently inaccessible or classified as proprietary. Consequently, river discharge measured at gauging stations downstream of the reservoir is often used as a proxy for reservoir operations. This approach inherently assumes that total reservoir outflow is equivalent to river discharge—an assumption that fails when significant volumes are diverted for hydropower, irrigation, or municipal supply. To overcome these limitations, satellite-derived products offer a promising alternative (Schwatke et al., 2015; Pekel et al., 2016; Schwatke et al., 2020; Donchyts et al., 2022; Khandelwal et al., 2022; Hao et al., 2024; Hou et al., 2024). These products offer the potential to monitor volume fluctuations in ungauged regions, making them crucial for future global model calibration and data assimilation (Hanazaki et al., 2022; Shen et al., 2025; Steyaert et al., 2025).

A calibration based on satellite products would target reservoir level or storage, instead of outflow (or river discharge as a proxy variable). To assess the feasibility of using satellite products, we compare the performance of the model when calibrated against three distinct targets: (i) reservoir release only, (ii) reservoir storage only, and (iii) a combined objective function utilizing both variables. By evaluating these configurations, this study aims to assess whether satellite-inferred storage can serve as a robust alternative to ground-based outflow records, thereby enabling improved reservoir modeling in ungauged or data-scarce regions.

The benchmarking is supported by the ResOpsUS dataset (Steyaert et al., 2022; Steyaert and Condon, 2024), the first comprehensive, national-scale inventory providing historical daily reservoir operations for 679 major dams across the CONUS. ResOpsUS includes records of reservoir inflow, level, storage and outflow (not all variables for all reservoirs), providing the necessary variables for our benchmark. We leverage this data-rich environment to test the models and calibration targets, establishing a proof-of-concept for operational environments where such granular data is unavailable.

In summary, this study benchmarks four reservoir modeling paradigms across the CONUS domain to address three primary objectives: (i) to identify the most robust reservoir routine for continental or global hydrological models; (ii) to determine the most suitable target variable for calibration—storage, outflow, or both; and (iii) to evaluate the feasibility of using remotely-sensed storage data in the calibration of large-scale models.

2 Data

2.1 Reservoir Selection and Filtering

To evaluate reservoir routines in a data-rich context, we utilized the ResOpsUS dataset (Steyaert et al., 2022; Steyaert and Condon, 2024), which provides daily time series of reservoir operations for 679 reservoirs across the CONUS. From this initial pool, we applied a rigorous filtering process to ensure the suitability of the reservoirs for LSHM:

1. Data Availability: We selected only those reservoirs containing concurrent daily records for all three primary variables: inflow (I), storage (V), and outflow (Q).

2. Physical Dimensions: Reservoirs were required to have a minimum catchment area of 50 km^2 and a minimum storage capacity ($V_{\max}$) of 10 hm^3.

3. Hydrological Impact: Following Shrestha et al. (2024), we excluded "non-disruptive" reservoirs—those that do not significantly alter the natural flow regime. This was quantified using the Degree of Regulation (DOR; Eq. 1a) and the Degree of Disruptivity (DOD; Eq. 1b). Reservoirs with DOR < 0.08 or DOD < 0.06 m were discarded:

$$\text{DOR} = \frac{V_{\max}}{\bar{I}_{\text{yr}}} \quad [-] \tag{1a}$$

$$\text{DOD} = \frac{V_{\max}}{A_c} \quad [\text{m}] \tag{1b}$$

where $V_{\max}$ is the reservoir storage capacity [m^3], $\bar{I}_{yr}$ is the average annual inflow [m^3/year], and A_c is the catchment area [m^2].

4. Record Length: To ensure sufficient data for robust model calibration, we removed reservoirs with time series shorter than four years.

5. Water Balance Integrity: Preliminary analysis revealed large biases between average inflow and outflow in several reservoirs. Such discrepancies may stem from poor data quality or significant consumptive use (e.g., water diversions) not explicitly represented in our modeling framework. To ensure a closed water balance for benchmarking, we removed reservoirs where the absolute bias exceeded 30%.

Applying these criteria resulted in a final selection of 164 reservoirs. Their geographical distribution, categorized by catchment area and storage capacity, is depicted in Fig. 1.

2.2 Dataset Integration and Attribute Enrichment

To satisfy the input requirements of the diverse reservoir routines, we enriched the ResOpsUS records with supplementary hydrometeorological time series, and catchment and reservoir attributes.

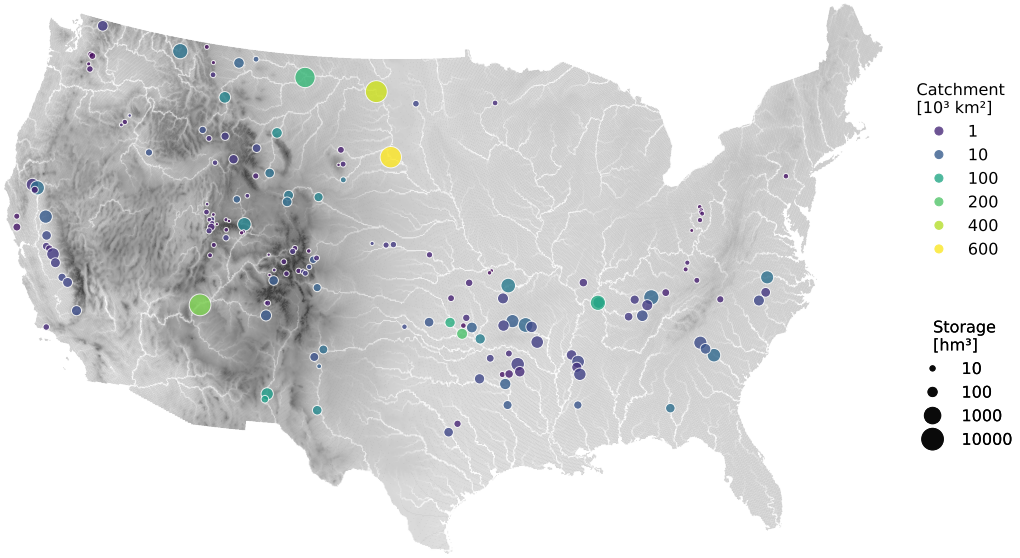

Figure 1. Distribution of the 164 reservoirs from the ResOpsUS dataset used in the benchmarking. Colors represent the catchment area and point size the storage capacity.

As detailed in Sect. 3.1, the routines require time series of direct precipitation and evaporation. These were extracted for each reservoir location from the ERA5 reanalysis dataset (Hersbach et al., 2020).

Static reservoir and dam attributes were primarily sourced from the Global Reservoir and Dam dataset (GRanD) (Lehner et al., 2011). While we included all GRanD attributes in our dataset, four were critical for the simulations: storage capacity ($V_{\max}$), maximum surface area ($A_{\max}$), dam crest elevation ($Z_{\max}$), and dam height (H_{dam}). A comparison of these attributes against observed time series revealed inconsistencies in some attribute values. To rectify this, we cross-referenced GRanD with the Global Dam Watch (GDW) dataset (Lehner et al., 2024). For instances where the attributes remained inconsistent, we manually updated the records using official data from U.S. authorities, including the U.S. Army Corps of Engineers' National Inventory of Dams (NID), the U.S. Bureau of Reclamation (USBR), and the U.S. Geological Survey (USGS) (US Army Corps of Engineers, 2025; US Bureau of Reclamation, 2025; US Geological Survey, 2025). Similar errors were reported by Shrestha et al. (2024).

For the purpose of parameter regionalization or the development of data-driven models, we further included catchment-scale characteristics derived from GloFAS surface fields (Choulga et al., 2024) and climate indices derived from ERA5 (Hersbach et al., 2020).

We named the enriched dataset *Reservoir Operations US and CAtchment and Reservoir Static attributes* (ResOpsUS+CARS). It is available via Zenodo. The dataset structure follows the standards defined by the CARAVAN initiative (Kratzert et al., 2023), ensuring it is formatted for immediate use in deep learning and large-sample hydrological studies.

Table 1. Data included in the ResOpsUS+CARS dataset used in this study.

Category	Variable	Source	Units	Purpose
Time	Inflow	[1]	$m^3 s^{-1}$	Model input
series	Storage	[1]	$10^6 m^3$	Calib./Eval.
	Outflow	[1]	$m^3 s^{-1}$	Calib./Eval.
	Elevation	[1]	m a.s.l.	Evaluation
	Precipitation	[2]	mm	Model input
	Evaporation	[2]	mm	Model input
Attributes	Capacity	[3]	$10^6 m^3$	Model input
	Surface Area	[3]	km^2	Area estim.
	Crest Elevation	[3]	m a.s.l.	Elev. estim.
	Dam Height	[3]	m	Elev. estim.

Sources: [1] ResOpsUS; [2] ERA5; [3] GRanD.

The specific time series and static attributes from ResOpsUS+CARS utilized in this study are summarized in Table 1. These include model forcing inputs, variables for calibration and evaluation, and geometric parameters used for the estimation of reservoir area or elevation.

3 Methods

3.1 Reservoir models

Each of the reservoir routines evaluated in this study is governed by the water mass balance equation (Eq. 2). At each time step Δt, the model calculates the reservoir storage at the end of the interval (V_{t+1}) based on the preceding storage state (V_t), the water surface area (A_t), and three forcing variables: inflow (I_t), precipitation (P_t), and evaporation (E_t):

$$V_{t+1} = V_t + (P_t - E_t) \cdot A_t + (I_t - Q_t) \cdot \Delta t \tag{2}$$

where Q_t is the reservoir outflow (release), the fundamental difference between the four reservoir schemes.

To represent the reservoir's bathymetry, we assume an inverted half-pyramid geometry (Fig. 2), following the approach of Liebe et al. (2005) and subsequent studies (Shin et al., 2019; Sadki et al., 2023; Shrestha et al., 2024).This geometric approximation allows for the continuous estimation of the water surface area (Eq. 3a), which is required for the mass balance, and the water level (Z_t; Eq. 3b), provided that the dam crest elevation and the dam height are known. We evaluate the validity of this geometric assumption by comparing estimated levels against observed records in a subset of reservoirs where such data is available.

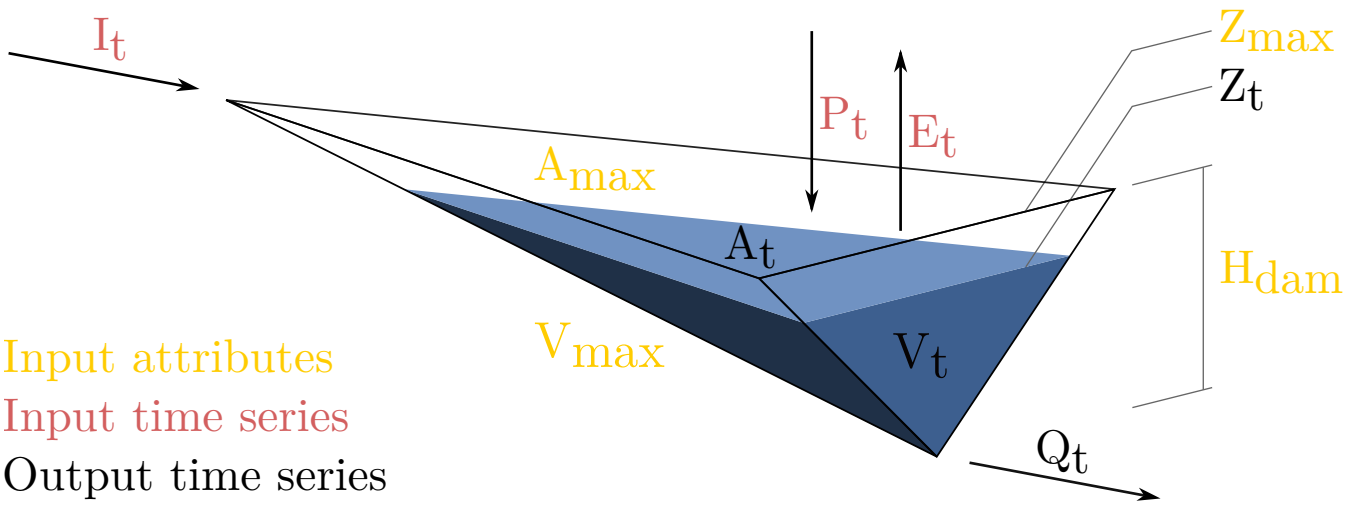

Figure 2. Simplified reservoir shape adopted in this study, including all the model input and output variables.

$$A_t = A_{\max}\left(\frac{V_t}{V_{\max}}\right)^{2/3} \tag{3a}$$

$$Z_t = Z_{\max} - H_{\text{dam}}\left[1 - \left(\frac{V_t}{V_{\max}}\right)^{1/3}\right] \tag{3b}$$

3.1.1 LISFLOOD

The current LISFLOOD reservoir routine (Burek et al., 2013; Zajac et al., 2017) operates as a piecewise linear storage-outflow relationship. The release is determined by the storage zone in which the reservoir resides at any given time step: conservative, normal or flood.

$$Q_t = \begin{cases} Q_c & \text{if } V_t < 2V_c \\ Q_c + (Q_n - Q_c)\frac{V_t - 2V_c}{V_n - 2V_c} & \text{if } 2V_c \le V_t < V_n \\ Q_n & \text{if } V_n \le V_t < V_n' \\ Q_n + (Q_f - Q_n)\frac{V_t - V_n'}{V_f - V_n'} & \text{if } V_n' \le V < V_f \\ \max\left(\frac{V_t - V_f}{\Delta t}, \min(Q_f, \max(kI_t, Q_n))\right) & \text{if } V_t > V_f \end{cases} \tag{4}$$

where V_c, V_n, V_n' and V_f are the conservative storage, the lower and upper bounds of the normal storage zone, and the flood storage limit, respectively. Q_c, Q_n and Q_f denote the outflow values associated with these limits. The release coefficient k

Table 2. Calibration parameters in the LISFLOOD reservoir module.

Parameter	Description	Definition	Minimum	Maximum	Default
α^{*}	Fraction of the total storage corresponding to the flood limit	$V_f = \alpha V_{\max}$	0.2	0.99	0.97
β	Proportion between flood limit and minimum storage corresponding to the normal limit	$V_n = V_c + \beta(V_f - V_c)$	0.001	0.999	0.655
γ^{**}	Proportion between flood and normal limits corresponding to the adjusted normal limit	$V_n' = V_n + \gamma(V_f - V_n)$	0.001	0.999	0.633
δ^{*}	Factor of the 100-year inflow (I_{100}) that defines the flood outflow	$Q_f = \delta I_{100}$	0.1	0.5	0.3
$\epsilon^{*,**}$	Ratio between normal and flood outflows	$Q_n = \epsilon Q_f$	0.001	0.999	$Q_f/\bar{I}$
k	Release coefficient		1	5	1.2

* Parameter identical in LISFLOOD and CaMa-Flood.

** Parameter adjusted in a standard LISFLOOD calibration.

allows for larger outflow than inflow when the reservoir exceeds the flood limit; while this was fixed at 1.2 in the original implementation, we treated it as a calibration parameter in this study (Table 2).

In the usual LISFLOOD calibration, only two reservoir parameters were tuned: the normal outflow (Q_n) and the upper limit of the normal storage zone (V_n'). The remaining breakpoints are kept at their default values to limit the dimensionality of the global calibration, which involves 14 parameters per basin.

However, in this benchmarking study, we allow maximum flexibility to the LISFLOOD routine by calibrating the six parameters listed in Table 2. This expanded parameterization serves two purposes. On one hand, it ensures a fair comparison with more flexible, data-driven schemes. On the other hand, because our decoupled calibration focuses exclusively on reservoir parameters rather than the entire hydrological model, we can afford the increased dimensionality to explore the routine's maximum performance potential.

A fundamental characteristic of this scheme is that it creates a unique (one-to-one) relationship between storage and outflow (Fig. 3). The outflow is strictly a function of storage, independent of seasonality, inflow or water demand. The only exception occurs in the flood zone ($V_t \geq V_f$), where the inclusion of the inflow (I_t) in the release logic allows for deviations from this static behavior. This makes the LISFLOOD scheme an ideal baseline for assessing whether more complex, dynamic routines are necessary to capture observed reservoir behavior.

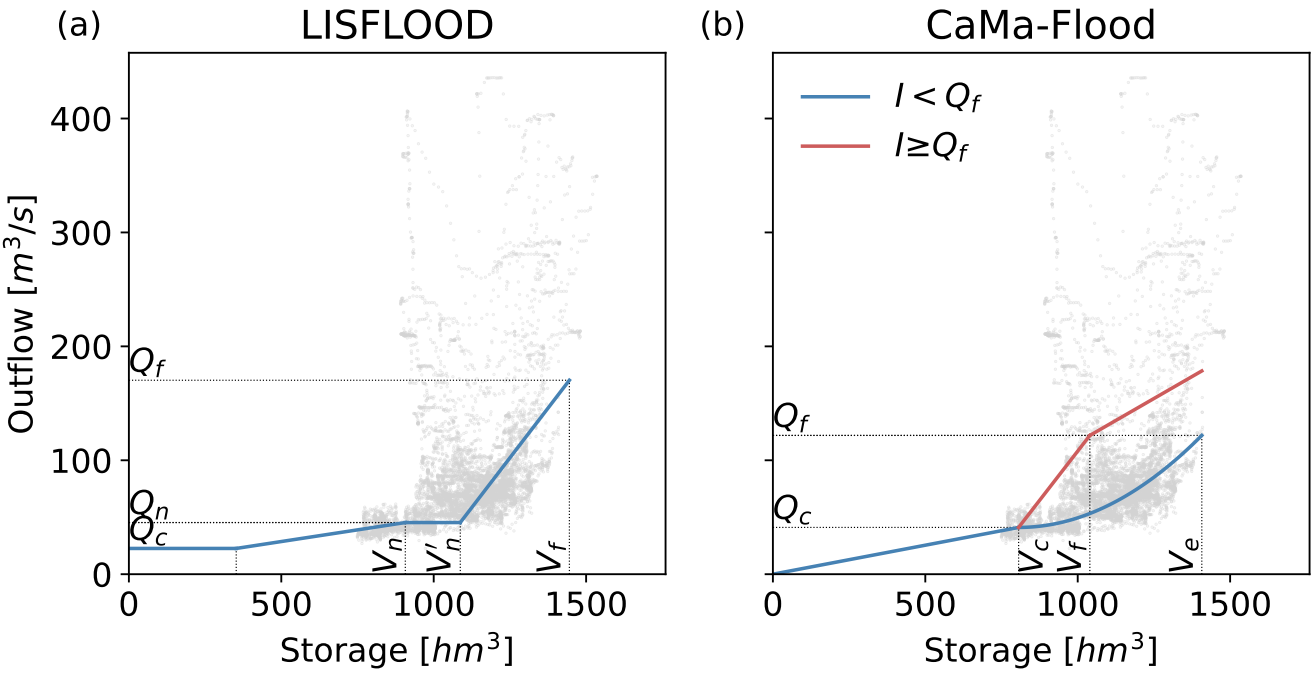

Figure 3. Storage–outflow relationships for the (a) LISFLOOD and (b) CaMa-Flood reservoir routines. The curves illustrate the calibrated operational rules for the Yellowtail dam (GRanD ID 355). Grey points represent observed daily records, highlighting the spread of historical operations.

3.1.2 CaMa-Flood

The reservoir module in the Catchment-based Macro-scale Floodplain model (CaMa-Flood) (Hanazaki et al., 2022) represents an evolution of the LISFLOOD approach by introducing inflow-dependency into the operational logic. Unlike the static LISFLOOD scheme, CaMa-Flood distinguishes between two operational modes based on the current inflow (I_t).

When I_t is below a threshold (Q_f), the outflow is modeled as a quadratic function of storage. This non-linear behavior restricts releases as the reservoir empties, effectively conserving water for future demand. Conversely, when I_t exceeds Q_f, the system transitions toward a linear reservoir behavior to manage high-flow conditions.

$$
Q_t = \begin{cases}
\begin{cases} Q_n \frac{V_t}{V_f} & V_t < V_c \quad \forall I_t \end{cases} \\
\begin{cases} Q_c + \left(\frac{V_t - V_c}{V_e - V_c}\right)^2 \Delta Q & V_c \le V_t < V_e \\ Q_f & V_t \ge V_e \end{cases} \quad I_t < Q_f \\
\begin{cases} Q_c + \frac{V_t - V_c}{V_f - V_c} \Delta Q & V_c \le V_t < V_f \\ Q_f + k \frac{V_t - V_f}{V_e - V_f} (I_t - Q_f) & V_f \le V_t < V_e \\ I_t & V_t \ge V_e \end{cases} \quad I_t \ge Q_f
\end{cases} \tag{5}
$$

where $Q_c = Q_n \dfrac{V_c}{V_f}$ and $\Delta Q = Q_f - Q_c$

where V_e is the emergency storage limit—defined as the upper part of the flood zone— and k is the release coefficient that regulates outflows under flood conditions based on available storage capacity. Hanazaki et al. (2022) defined k as a fixed value derived from the total flood volume (Eq. 6a). Under that formulation, k becomes zero for high-regulation reservoirs, resulting in a constant release even as storage levels approach V_e, effectively neglecting the exhaustion of regulation capacity. To address this unrealistic behavior, we introduced a dynamic calculation of k at each time step based on the instantaneous reservoir volume (V_t, Eq. 6b). This modification ensures that k increases as storage approaches V_e, facilitating higher discharge to mitigate the risk of overtopping.

$$k = \max\left(1 - \frac{V_{\max} - V_f}{A_c \cdot 0.2}, 0\right) \tag{6a}$$

$$k = \max\left(1 - \frac{V_{\max} - V_t}{A_c \cdot 0.2}, 0\right) \tag{6b}$$

While Hanazaki et al. (2022) utilized default parameters, we calibrated five parameters (Table 3) to ensure the routine has comparable degrees of freedom to the other routines. To maintain consistency with the LISFLOOD benchmark, three parameters (α, δ, ϵ) share identical definition across both models.

The primary advantage of the CaMa-Flood routine over LISFLOOD is its ability to represent a non-unique relationship between storage and outflow (Fig. 3). By incorporating I_t into the operational rules, the routine can better capture the observed dispersion in the storage-outflow values, providing a more realistic representation of human-regulated flow variability.

3.1.3 mHM

The reservoir module within the Multiscale Hydrological Model (mHM) (Samaniego et al., 2010; Kumar et al., 2013; Thober et al., 2019) represents a fundamental shift in modeling philosophy compared to the LISFLOOD and CaMa-Flood routines. Rather than treating outflow as a direct function of storage, the mHM routine determines releases primarily based on a water

Table 3. Calibration parameters in the CaMa-Flood reservoir module.

Parameter	Description	Definition	Minimum	Maximum	Default
α^*	Fraction of the total storage corresponding to the flood limit	$V_f = \alpha V_{\max}$	0.2	0.99	0.97
β	Proportion between flood limit and total storage corresponding to the extreme limit	$V_e = V_{\max} - \beta(V_{\max} - V_f)$	0.001	0.999	0.2
γ	Proportion of the flood limit corresponding to the normal limit	$V_c = \gamma V_f$	0.001	0.999	0.5
δ^*	Factor of the 100-year inflow (I_{100}) that defines the flood outflow	$Q_f = \delta I_{100}$	0.1	0.5	0.3
ϵ^*	Ratio between normal and flood outflows	$Q_n = \epsilon Q_f$	0.001	0.999	$Q_f/\bar{I}$

* Parameter identical in LISFLOOD and CaMa-Flood.

demand time series (D_t). While Shrestha et al. (2024) utilized random forests to predict D_t for individual reservoirs, we implement a seasonal approach to maintain model portability.

In our implementation, we derive an annual demand climatology by computing the mean observed outflow for each day of the year across the available records. To ensure this signal reflects demand rather than total outflow, we apply the water stress ($\bar{D}/\bar{I}$) as scaling factor, and smooth the signal using a 28-day moving average. This smoothed climatology serves as the seasonal demand signal D_t for each reservoir, introducing the temporal variability missing in static storage–discharge schemes.

Reservoir release. The total daily release Q_t is partitioned into a component satisfying the hedged demand ($\hat{D}_t$) and a component representing the passthrough of inflow (I_t):

$$Q_t = \rho \cdot \kappa_t \cdot \hat{D}_t + (1 - \rho) \cdot I_t \tag{7}$$

The partition coefficient ρ is a function of the reservoir's degree of regulation (DOR). For highly regulated reservoirs (DOR $\geq \alpha$), the release is entirely demand-driven ($\rho = 1$). In all other cases, the influence of demand relative to inflow is governed by the parameters α and β:

$$\rho = \min\left(1, \left(\frac{\text{DOR}}{\alpha}\right)^{\beta}\right) \tag{8}$$

The demand component is further moderated by a time-varying coefficient κ_t, which hedges the release based on the current storage level V_t relative to a target normal storage V_n:

$$\kappa_t = \left(\frac{V_t}{V_n}\right)^{\lambda} = \left(\frac{V_t}{\gamma \cdot V_{\max}}\right)^{\lambda} \tag{9}$$

Table 4. Calibration parameters in the mHM reservoir module.

Parameter	Description	Definition	Minimum	Maximum	Default
α	Threshold in the degree of regulation that defines demand-controlled reservoirs (DOR $> \alpha$)	Eq. 8	0.0	5.0	0.5
β	It controls indirectly the proportion of inflow and demand in the releases	Eq. 8	0.5	3.0	1.0
γ	Ration between normal and total storage	Eq. 9	0.0	1.0	0.85
λ	It further controls hedging based on the current reservoir storage	Eq. 9	0.25	3.0	1.0
ω	It controls hedging based on the ratio between the current and average demand	Eq. 10	0.0	1.0	0.1

where γ defines the normal storage fraction and λ controls the sensitivity of the hedging to storage deficits.

Demand hedging. The hedged demand $\hat{D}_t$ is calculated based on the ratio between average annual demand ($\overline{D}$) and inflow ($\overline{I}$). The parameter ω represents the expected annual water surplus and is used to categorize the reservoir's water stress. If there is no water stress ($\overline{D}/\overline{I} < 1-\omega$), the daily demand ($D_t$) is supplied in addition to the average water surplus. If there is stress, the hedged demand is a combination of a fraction of the average inflow and a fraction of the daily demand.

$$\hat{D}_t = \begin{cases} \overline{I} - \overline{D} + D_t & \text{if } \frac{\overline{D}}{\overline{I}} < 1-\omega \\ \omega \cdot \overline{I} + (1-\omega)\frac{\overline{I}}{\overline{D}} \cdot D_t & \text{otherwise} \end{cases} \tag{10}$$

Calibration. Following the procedure established by Shrestha et al. (2024), we calibrated five parameters (Table 4). Other authors used default values of some of the parameters (Hanasaki et al., 2006; Shin et al., 2019; Sadki et al., 2023), which we have used for the simulation with default parameters.

Unlike the rule-based schemes, the mHM routine produces a more realistic storage–outflow relationship, as the release is decoupled from the instantaneous storage volume and instead follows the seasonal cycle of demand. While more flexible, the main limitation of the routine is how to estimate the demand signals.

3.1.4 STARFIT

The Storage Targets And Release Function Inference Tool (STARFIT) framework represents a data-driven paradigm that infers operational rules by learning the seasonality of reservoir storage and outflows (Turner et al., 2021). Unlike previous rule-based models, STARFIT employs harmonic functions to represent seasonal cycles. A key advantage of this formulation is its scale-independence; while the model parameters are typically fitted using weekly data to reduce noise, the harmonic frequency (ω) can be adjusted for daily simulations ($\omega = 1/365$), ensuring consistency with the temporal resolution of the simulation.

All variables are standardized to facilitate regionalization and comparison across basins (Eq. 11). Storage is converted to reservoir filling, while inflow and outflow are expressed as normalized anomalies relative to the mean annual inflow ($\bar{I}$):

$$\hat{V}_t = \frac{V_t}{V_{\max}} \tag{11a}$$

$$\hat{Q}_t = \frac{Q_t - \bar{I}}{\bar{I}} \tag{11b}$$

$$\hat{I}_t = \frac{I_t - \bar{I}}{\bar{I}} \tag{11c}$$

Storage Normal Operating Range (NOR). STARFIT defines the operational target of a reservoir through a Normal Operating Range (NOR), bounded by upper (NOR_{up}) and lower (NOR_{low}) seasonal harmonics. These bounds represent the typical filling levels for any given time of the year:

$$\begin{aligned} \text{NOR}_{\text{up},t} &= \min\Big(\max\big(A + B\sin 2\pi\omega t + C\cos 2\pi\omega t,\ \hat{V}_{\min}\big), \\ &\qquad \hat{V}_{\max}\Big) \\ \text{NOR}_{\text{low},t} &= \min\Big(\max\big(a + b\sin 2\pi\omega t + c\cos 2\pi\omega t,\ \hat{v}_{\min}\big), \\ &\qquad \hat{v}_{\max}\Big) \end{aligned} \tag{12}$$

where t is the time (day or week) of the year. The NOR has 10 parameters: six defining the harmonic curves and four capping the maximum and minimum filling. To fit these curves, the model uses only the three most extreme (highest and lowest) observed storage for each calendar week (Fig. 4), ensuring the NOR captures the envelope of historical operations.

Release function. The release policy is conditioned on the reservoir's position relative to the NOR. To maintain the reservoir within its seasonal bounds, STARFIT applies a three-zone operational policy:

$$Q_t = \begin{cases} \min\Big(Q_{\min} + (Q_{\text{NOR},t} - Q_{\min})\frac{\hat{V}_t}{\text{NOR}_{\text{low},t}}, I_t\Big) & \text{if } \hat{V}_t < \text{NOR}_{\text{low},t} \\ \max\Big(Q_{\min}, \min(Q_{\text{NOR},t}, Q_{\max})\Big) & \text{if } \hat{V}_t \in \text{NOR}_t \\ Q_{\text{NOR},t} + (Q_{\max} - Q_{\text{NOR},t})\frac{\hat{V}_t - \text{NOR}_{\text{up},t}}{1 - \text{NOR}_{\text{up},t}} & \text{if } \hat{V}_t > \text{NOR}_{\text{up},t} \end{cases} \tag{13}$$

We modified the definitions of the outflow for both below and above the NOR, as the original definitions produced in some cases noisy releases. In both cases, we have introduced a linear interpolation between a fixed minium or maximum outflow ($Q_{\min}$ or $Q_{\max}$) and the outflow associated with the respective NOR boundary ($Q_{\text{NOR}_{\text{low}},t}$ or $Q_{\text{NOR}_{\text{up}},t}$). Within the NOR, the standardized outflow ($\hat{Q}_t$) is modeled as the sum of an harmonic component ($\hat{Q}_{harm,t}$) and a linear component ($\hat{Q}_{\text{lin},t}$) that accounts for the current storage and inflow:

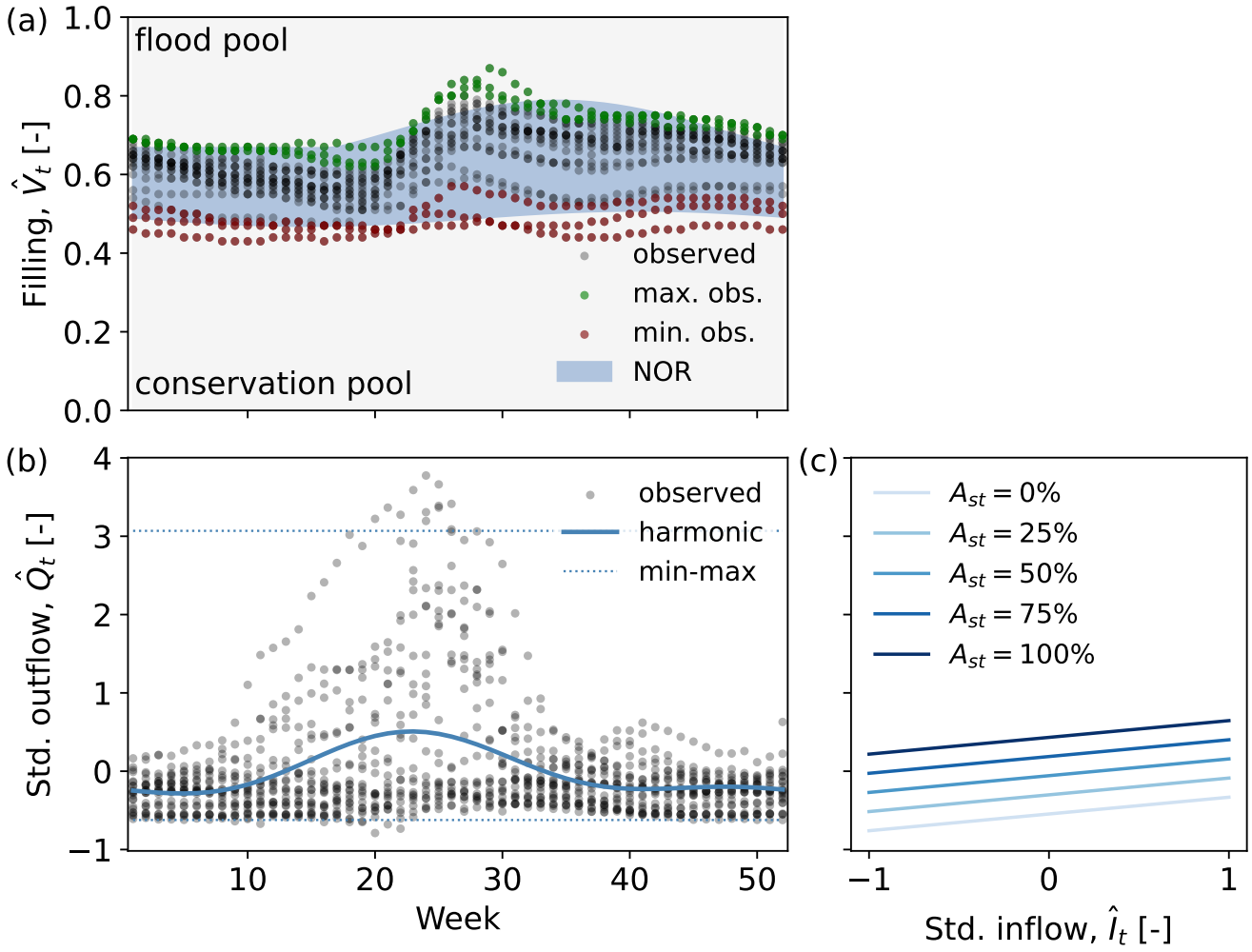

Figure 4. The STARFIT model for the Yellowtail dam (GRanD ID 355): (a) seasonal harmonic curves defining the storage normal operating range (NOR) for each calendar week; (b) the seasonal harmonic release; (c) the linear model of release residuals based on available storage (A_{st}) and standardized inflow.

$$\hat{Q}_{\text{NOR},t} = \hat{Q}_{\text{harm},t} + \hat{Q}_{\text{lin},t}$$

$$\hat{Q}_{\text{harm},t} = d \sin 2\pi\omega t + e \cos 2\pi\omega t + f \sin 4\pi\omega t + g \cos 4\pi\omega t \tag{14}$$

$$\hat{Q}_{\text{lin},t} = h + i \frac{\hat{V}_t - \text{NOR}_{\text{low},t}}{\text{NOR}_{\text{up},t} - \text{NOR}_{\text{low},t}} + j \cdot \hat{I}_t$$

The release harmonic allows for bimodal release patterns (e.g., two release peaks per year). The linear term is discarded if the linearity is weak ($R^2 < 0.2$), reverting to a purely seasonal policy.

The complete release model comprises 7 parameters, fitted using the full historical record of weekly releases (Fig. 4). Additionally, the minimum and maximum releases ($Q_{\min}$ and $Q_{\max}$) are estimated as quantiles of the observed daily inflows; we have changed the original values of these quantiles to 1% and 99%, respectively.

3.2 Experimental design

To address our research questions, we conducted four distinct simulation experiments designed to evaluate the routines under varying levels of data availability and calibration complexity.

The first run evaluates model performance using default parameters sourced from the literature (see Tables 2, 3, and 4). This experiment assesses the "out-of-the-box" reliability of the routines, which is particularly relevant for continental or global applications where historical operational data are unavailable.

The remaining three experiments involve local parameter optimization targeting different target variables:

- OBJ-Q: Univariate calibration of reservoir outflow. This mirrors the standard calibration procedure for operational systems like EFAS/GloFAS, where model parameters are tuned to match observed streamflow.
- OBJ-V: Univariate calibration of reservoir storage. This experiment evaluates the potential for training routines using satellite-derived storage products, which is vital for ungauged basins.
- OBJ-QV: Bivariate calibration of both storage and outflow. This multi-objective approach analyzes the trade-offs in model fidelity and identifies the information loss associated with single-variable calibration.

Parameter optimization was performed using the Shuffled Complex Evolution (SCEUA) algorithm (Duan et al., 1993, 1994), implemented via the *Spotpy* Python library (Houska et al., 2015). For each reservoir, the algorithm was executed for up to 1,000 iterations using four complexes. Following the recommendations of Shen et al. (2022) for large-sample hydrology, we used the full length of the available daily time series for calibration to capture the widest possible range of hydrological conditions.

Performance was quantified using the modified Kling-Gupta Efficiency (KGE′) (Kling et al., 2012), which decomposes performance into correlation, bias, and variability components. For the bivariate experiment, we integrated the individual KGE′ scores into a single multi-objective metric:

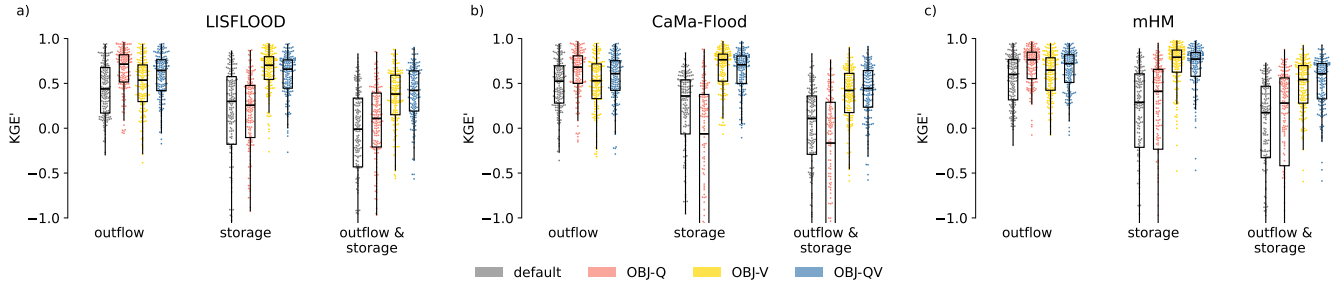

Figure 5. Comparison of the model performance depending on the variable(s) targeted in the calibration. The panels corresponds to the three reservoirs routines that were calibrated: a) LISFLOOD, b) CaMa-Flood and c) mHM. Within each panel, three groups of box plots show the performance in terms of outflow, storage, and both variables together. Four parameter sets are compared —default values, calibration of outflow (OBJ-Q), storage (OBJ-V) and both variables (OBJ-QV)— indicated by colors. The dots represent individual reservoirs, indicating the distribution of the overall performance.

$$\mathrm{KGE}' = 1 - \sqrt{(r-1)^2 + \left(\frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mathrm{CV}_{\mathrm{sim}}}{\mathrm{CV}_{\mathrm{obs}}} - 1\right)^2} \tag{15a}$$

$$\mathrm{KGE}'_{\mathrm{bivariate}} = 1 - \sqrt{(1-\mathrm{KGE}'_Q)^2 + (1-\mathrm{KGE}'_V)^2} \tag{15b}$$

where r is the Pearson correlation coefficient, μ is the mean, and CV is the coefficient of variation.

While the LISFLOOD, CaMa-Flood, and mHM routines are optimized using the iterative SCEUA algorithm, STARFIT employs a direct statistical fitting procedure. This approach does not require a heuristic search. Instead, the storage model parameters are estimated in a two-step process: an initial ordinary least squares (OLS) regression provides a first guess for the three harmonic parameters, followed by a non-linear optimization using the L-BFGS-B algorithm to finalize all five parameters, including the upper and lower capping values ($\hat{V}_{\max}, \hat{V}_{\min}$). In contrast, the parameters for the release model are derived exclusively through OLS regression.

Because the original STARFIT formulation inherently requires observations of both storage (to define the NOR) and outflow (to define the release policy), it is only evaluated within the OBJ-QV bivariate experiment. This ensures a consistent comparison between the statistically-derived STARFIT policies and the calibrated rule-based routines.

4 Results

4.1 Which target variable is most relevant for reservoir calibration?

Figure 5 illustrates the performance of the three rule-based reservoir routines (LISFLOOD, CaMa-Flood, and mHM) across the four experimental configurations. The results highlight a critical dependency between the calibration objective and the model's ability to represent the dual states of discharge and storage.

The default parameterization (gray) provides a reasonably effective baseline for reservoir outflow across all three routines, with median KGE′ values of 0.44, 0.52, and 0.60 for LISFLOOD, CaMa-Flood, and mHM, respectively. However, these default parameters fail to capture storage dynamics effectively; median KGE′ scores for storage drop to 0.30, 0.36, and 0.29, respectively. The significantly larger interquartile range (IQR) in these scores indicates the high variability in storage performance across the 164 reservoirs.

As expected, each routine achieves its peak performance for a specific variable when that variable is the sole target of the calibration. However, there is a clear trade-off when we look at how the models perform on variables they weren't trained on. Calibrating outflow (OBJ-Q) severely degrades storage performance; in LISFLOOD and CaMa-Flood, this configuration yields the lowest median KGE′ for storage among all experiments; in mHM, while the median is slightly higher than the default parameters, the dispersion between reservoirs is remarkably large. Conversely, calibrating to storage does not result in a similar degradation of outflow performance. In all three routines, OBJ-V maintains outflow KGE′ values that are superior to the default parameterization.

The bivariate calibration (OBJ-QV) consistently achieves the highest combined performance across all routines. Notably, the performance of the OBJ-QV run is closely mirrored by the storage-only calibration (OBJ-V). These findings suggest that reservoir storage is a more informative variable than outflow for identifying model parameters.

While operational hydrology has traditionally relied on streamflow (outflow) for calibration, our results indicate that storage time series provide a more robust constraint on the internal state of the reservoir. This has significant implications for global hydrology, as it supports the feasibility of using satellite-derived storage products to calibrate reservoir routines in ungauged or data-poor regions (Shen et al., 2025).

4.2 Which is the best-performing reservoir routine?

Figure 6 compares the performance of the four routines across the different experimental configurations. Note that the outflow-only calibration (OBJ-Q) is excluded here for simplicity, as the previous section demonstrated that it is a suboptimal strategy for capturing overall reservoir behavior.

Among the rule-based schemes, the mHM routine consistently achieves the highest performance across nearly all scenarios. Its primary strength lies in its ability to represent seasonal operations through demand-driven logic. The only exception is the uncalibrated (default) simulation of storage, where CaMa-Flood shows a slight advantage. Note that the CaMa-Flood routine was specifically designed for global applications without site-specific tuning (Hanazaki et al., 2022).

The results also highlight the benefits of model complexity: the CaMa-Flood routine, which evolved from the original LISFLOOD scheme by adding quadratic storage logic and inflow-dependency, systematically outperforms LISFLOOD in every experiment. This suggests that even small increases in the physical complexity of the storage-discharge relationship can significantly improve model performance.

Finally, the bivariate calibration results provide the first direct comparison with the STARFIT model. Despite its high parameterization (up to 19 parameters) and seasonal flexibility, STARFIT does not outperform mHM. Its combined performance

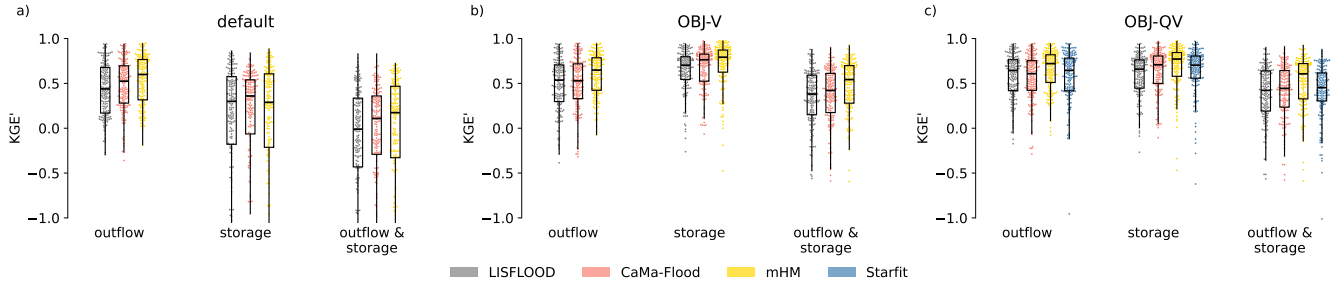

Figure 6. Comparison of the model performance depending on the reservoir routine. The panels corresponds to three of the experiments: a) default values, b) calibration of storage (OBJ-V), c) bivariate calibration of outflow and storage (OBJ-QV). Within each panel, three groups of box plots show the performance in terms of outflow, storage, and both variables together. Four reservoir routines are compared— LISFLOOD, CaMa-Flood, mHM and STARFIT— indicated by colors. The dots represent individual reservoirs, indicating the distribution of overall performance.

is generally comparable to CaMa-Flood. This suggests that while seasonal harmonics are valuable, the structured rules in mHM and CaMa-Flood may provide a more robust framework for continuous daily simulations in large-scale hydrological models.

4.3 How accurate is the reservoir shape approximation?

To represent the elevation–storage–area relationship in our modeling, we adopted a simplified geometric assumption where all reservoirs are treated as an inverted half-pyramid (Liebe et al., 2005). Under this assumption, the reservoir geometry is defined by the maximum surface area ($A_{\max}$) and the dam height (H_{dam}) (Eq. 3). This approximation directly influences the calculation of the instantaneous surface area, which is a critical term for estimating direct precipitation and open-water evaporation.

While in situ observations of reservoir area are often unavailable, we validated this geometric assumption by comparing time series of estimated and observed water elevation. Time series of observed storage were converted into water elevations via Eq. (3b). These estimates were then compared against in situ observations for a subset of 138 reservoirs where such data were available. The performance of this elevation estimation is summarized in Fig. 7.

The inverted half-pyramid assumption proved to be a robust approximation for the majority of the study sites. The median KGE′ for reservoir level estimation is 0.72, with 75% of the reservoirs achieving a KGE′ at least 0.44. From the three components of KGE′, the variability is the main cause of poor perfomance (median value of 1.25), whereas bias and correlation showed an overall good performance (median values of 1.00 in both cases). The accuracy of the level calculation is highly sensitive to the quality of the static attributes, specifically the crest elevation ($Z_{\max}$) and dam height (H_{dam}) (Eq. 3b). As discussed in Section 2.2, errors in global databases such as GRanD or GDW can propagate into these elevation estimates; therefore, data cleaning and the use of locally validated attributes are essential for accurate water-level modeling.

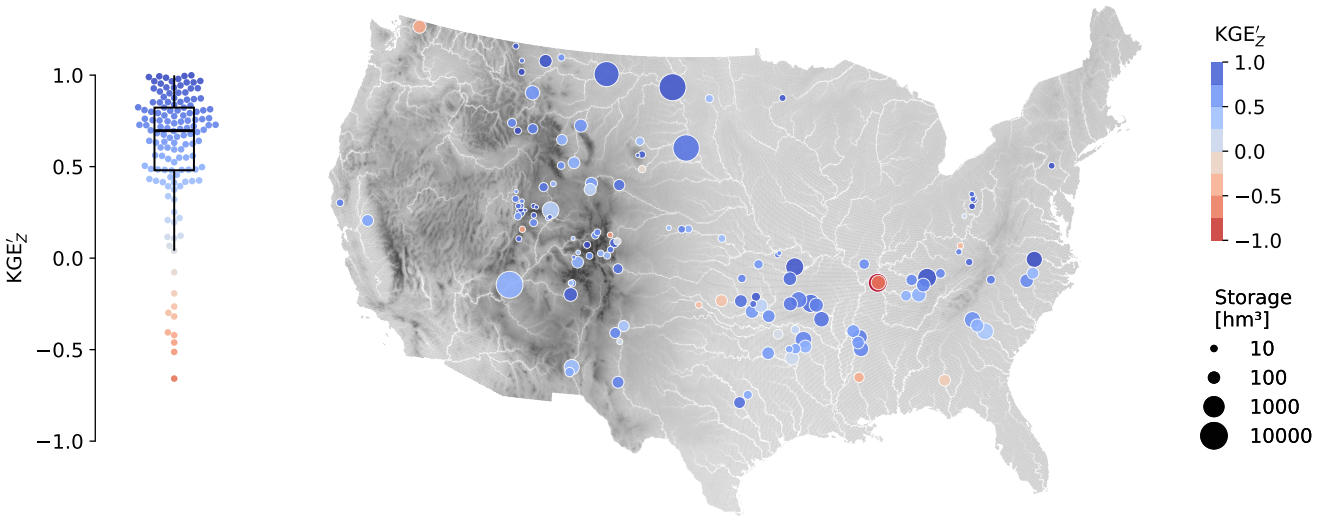

Figure 7. Performance of the reservoir level estimation assuming an inverted half-pyramid reservoir shape. The swarm plot on the left shows the overall distribution, where each point represents a reservoir and the box plot the quartiles. The map on the right shows the geographical distribution of performance; the dot size indicates storage capacity.

5 Discussion

5.1 Balancing model complexity and operational feasibility

This study compared reservoir routines of varying complexity, from the simple storage-dependent logic of LISFLOOD to the demand-driven and statistically-inferred policies of mHM and STARFIT. Our results indicate that model performance generally scales with the number of environmental covariates included in the routine (e.g., inflow or demand).

While mHM emerged as the best-performing routine (Section 4.2), its dependence on a demand time series poses a significant operational challenge. In data-rich environments, demand can be inferred via machine learning (Shrestha et al., 2024), but in continental or global systems like EFAS or GloFAS, such data are often unavailable.

A similar constraint applies to STARFIT, although its performance was not markedly superior to the more parsimonious CaMa-Flood routine. A distinct advantage of STARFIT for global applications is its flexibility across different temporal resolutions. For instance, Steyaert et al. (2025) recently demonstrated a global application of the STARFIT routine within the PCR-GLOBWB2 hydrological model by leveraging satellite-derived storage from the GloLAKES dataset (Hou et al., 2024) to define the storage normal operating range (NOR).

The CaMa-Flood routine has been selected for implementation in the forthcoming major evolution of the open-source LISFLOOD model (OS-LISFLOOD). This update will form the core of the upcoming EFAS v6 and GloFAS v5 operational systems (Salamon et al., 2025). These systems undergo regular updates encompassing updated meteorological forcing, revised

surface fields, extended river discharge records and model developments. By evolving the reservoir logic to include inflow-dependency and quadratic storage–outflow relationships, we achieve a measurable gain in performance without increasing the data requirements for calibration. Furthermore, the robust default performance of the CaMa-Flood routine supports global applications in regions where historical records are sparse (Hanazaki et al., 2022).

5.2 Current limitations and the forecast challenge

A notable limitation across the benchmarked routines is the absence of explicit reservoir purpose (e.g., irrigation, hydropower, or flood control...) (Shen et al., 2025). Reservoir operation is fundamentally driven by water demand, which varies significantly by use case. For instance, hydropower reservoirs are often maintained at high levels to maximize hydraulic head, whereas flood control reservoirs are drawn down ahead of wet seasons to maximize storage capacity. Conversely, irrigation reservoirs follow a distinct seasonal filling and release cycle to bridge the gap between wet-season supply and dry-season demand.

While simple schemes like LISFLOOD or CaMa-Flood do not explicitly represent these operational rules, several studies have focused on purpose-specific modeling. This includes routines for water supply in the UK (Salwey et al., 2024), hydropower in France (Baratgin et al., 2024), and non-consumptive uses (Shrestha et al., 2024). Other approaches differentiate between irrigation and non-irrigation schemes (Hanasaki et al., 2006; Sadki et al., 2023) or apply distinct objective functions based on the reservoir's primary use (Haddeland et al., 2006).

In our comparison, the superior performance of mHM and STARFIT likely stems from their ability to capture these dynamics indirectly. mHM is a demand-driven routine that can infer specific usage patterns directly from demand time series, while STARFIT leverages observed seasonal patterns to learn operational behaviors. The fact that these two models outperformed LISFLOOD and CaMa-Flood, even if by a small margin, suggests that accounting for the temporal variability of human-driven demand is critical for improving global reservoir modeling.

Furthermore, these routines lack a forward-looking component, whereas in practice, reservoir operators rely on meteorological and hydrological forecasts across multiple timescales. While replacing current inflow with forecasted values could address this, it introduces a significant bottleneck for computationally demanding systems. The model would need to simulate the entire upstream catchment for the full forecast horizon before the reservoir release can be determined at any given node. While local or regional systems running simpler lumped models could potentially process these computations sequentially at each reservoir location, this remains a substantial implementation challenge for continental or global distributed systems.

5.3 The potential of satellite products

The scarcity of in situ records remains the primary hurdle for large-scale reservoir modeling (Shen et al., 2025). Satellite-derived products offering reservoir area, elevation, and storage anomalies provide an opportunity to bridge this gap (Zhao and Gao, 2018; Schwatke et al., 2020; Donchyts et al., 2022; Khandelwal et al., 2022; Hao et al., 2024; Hou et al., 2024). Two of the outcomes of this study specifically support the transition toward using satellite products. First, we found that reservoir storage is a more informative calibration target than reservoir outflow, as storage-based calibration yields models that perform well across both state variables (Section 4.1). This represents a "lucky coincidence" for the community, as many available satellite products

provide direct estimates of reservoir storage or storage variations rather than discharge. Second, our validation of the inverted half-pyramid shape (Section 4.3) suggests that even simplified geometric assumptions are sufficient to translate satellite-derived levels into volumes with reasonable accuracy. However, a rigorous analysis of the feasibility of remotely sensed data should consider also spatial and temporal , apart from accuracy.

Recent studies are already leveraging these synergies; for instance, Shen et al. (2025) estimated monthly target storage values for 1,178 GRanD reservoirs by combining area time series from Global Water Watch (GWW) (Donchyts et al., 2022) with area-storage curves from GRDL (Hao et al., 2024). This monthly target storage was integrated into an evolved reservoir scheme—based on LISFLOOD and CaMa-Flood—to introduce seasonality into otherwise simple storage–outflow formulations.

The flexibility of STARFIT's harmonic formulation is particularly well-suited for satellite data, as emphasized by Steyaert et al. (2025). Because the frequency term in the harmonic curves allows the model to be fitted at coarser temporal resolutions (e.g., monthly satellite passes) and then run at a daily time step, it overcomes the temporal sampling limitations of many satellite products.

In line with Shen et al. (2025), we have conducted a preliminary analysis of the potential use of the (GWW) dataset (Donchyts et al., 2022), as detailed in Appendix A. Our findings indicate that while satellite-derived storage follows observed historical trends, a persistent bias often exists between in situ and remotely-sensed values. The primary bottleneck for the quality of these products is the availability of accurate elevation-storage-area curves. While several efforts have attempted to provide these curves globally (Yigzaw et al., 2018; Busker et al., 2019; Khazaei et al., 2022), they frequently show shortcomings when validated against in situ data.

5.4 Toward data-driven reservoir modeling

The recent revolution in deep learning has demonstrated that data-driven models can outperform traditional physical models in streamflow prediction (Nearing et al., 2024). To support a similar transition in reservoir modeling, we have formatted the ResOpsUS+CARS dataset following the CARAVAN standard (Kratzert et al., 2023). This extension of the original ResOpsUS dataset (Steyaert et al., 2022) integrates the reservoir operations with meteorological time series, the reservoir attributes from GRanD/GDW, catchment characteristics from GloFAS static maps, and climate indices from ERA5. By consolidating these diverse covariates—encompassing reservoir morphology, catchment physiography, seasonality, meteorology and climate—we provide the necessary information for deep learning models to identify the complex drivers of reservoir release.

Reservoir operations are the product of complex human-natural interactions that process-based routines struggle to capture entirely. While we acknowledge that deep learning offers a path toward superior performance, these models are notoriously data-hungry. This highlights the ongoing necessity of curated, multi-national observational datasets—not just for training, but for the rigorous benchmarking of the hybrid physical-statistical models and the development of global remotely-sensed approaches.

6 Conclusions

This study provided a comprehensive benchmarking of four reservoir routines (LISFLOOD, CaMa-Flood, mHM, and STARFIT) evaluated across 164 reservoirs in the United States using the ResOpsUS and GRanD datasets. By comparing different calibration strategies and model structures, we draw the following conclusions.

Reservoir storage is the most informative variable for calibration. Our results demonstrate a significant asymmetry in model identifiability. While calibrating routines to match observed outflow (the standard practice in many hydrological models) leads to a poor representation of internal reservoir states, calibrating to observed storage yields robust performance for both storage and outflow. This finding supports the shift toward using satellite-derived storage products for model parameterization, particularly in ungauged basins.

mHM offers superior performance, but CaMa-Flood provides the best operational balance. While the demand-driven logic of the mHM routine consistently demonstrated superior performance, its reliance on site-specific demand data constrains its applicability in continental and global systems where such information is often scarce. The CaMa-Flood routine, which will be integrated into the OS-LISFLOOD model, represents an optimal compromise: it provides significant improvements over the original linear storage logic without requiring additional input variables or data-intensive calibration.

The inverted half-pyramid geometric assumption is a robust approximation. Validating reservoir level estimations against in situ observations confirmed that a simplified inverted half-pyramid shape is sufficient for large-scale modeling. This validates the use of maximum area and dam height as sufficient proxies to link storage, area, and elevation, which is critical for estimating evaporative losses and integrating altimetry or reservoir area data.

STARFIT's flexibility is a key asset for remote sensing integration. Although STARFIT did not outperform the rule-based mHM or CaMa-Flood routines in daily simulations, its harmonic formulation is uniquely capable of bridging temporal gaps in observations. Its ability to be fitted at monthly scales and run at daily resolutions makes it a powerful tool for leveraging current and future satellite altimetry missions.

Looking forward, the integration of reservoir modeling into global hydrology must move beyond the "data-poor" paradigm. The creation of standardized, multi-national datasets like the ResOpsUS+CARS is a step toward this goal. By providing these data in the CARAVAN format, we aim to facilitate the development of hybrid models that combine the physical consistency of process-based routines with the predictive power of deep learning. Such advancements will be essential for improving the accuracy of operational flood and drought forecasting systems in an increasingly managed global water cycle.

Code and data availability. All the code used to generate the reservoir dataset, implement the four reservoir schemes, and perform the model calibrations is available in the following GitHub repository: https://github.com/casadoj/reservoirs-LSHM. The benchmarking dataset generated for this study can be accessed via Zenodo at https://doi.org/10.5281/zenodo.15978041.

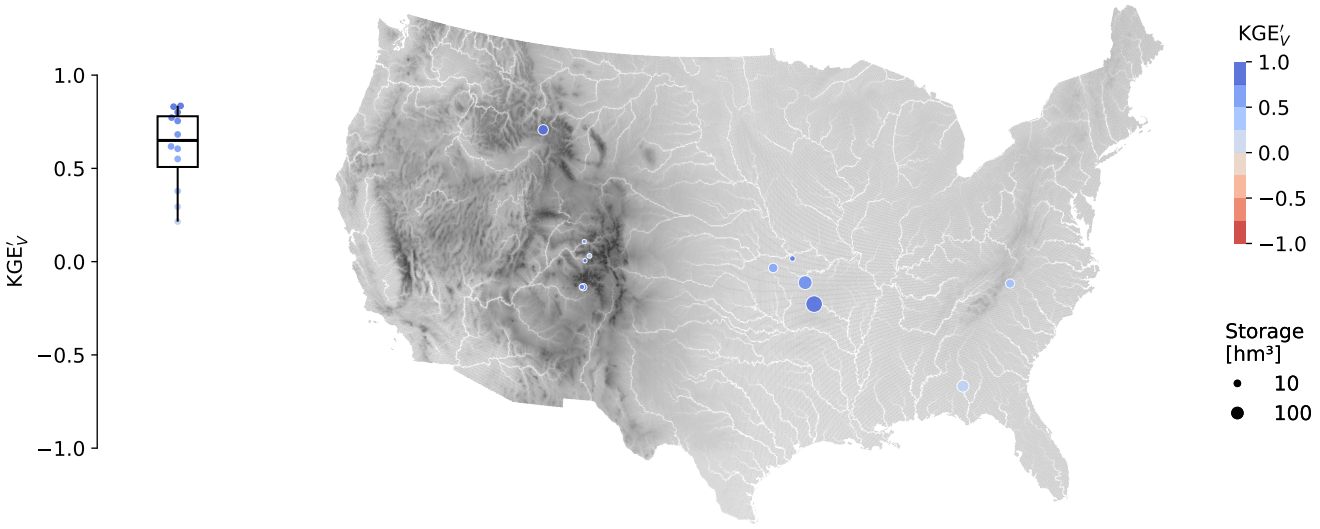

Figure A1. Performance of the reservoir storage estimates provided by the Global Water Watch. The swarm plot on the left shows the overall distribution, where each point represents a reservoir and the box plot the quartiles. The map on the right shows the geographical distribution of performance; the dot size indicates storage capacity.

Appendix A: Evaluatin of GWW storage estimates

To assess the feasibility of using Global Water Watch (GWW) (Donchyts et al., 2022) for future global-scale reservoir calibration, we evaluated the accuracy of its volume estimates against our observational sample. Although GWW provides reservoir area for nearly all reservoirs in our dataset, storage data were available for only 12 of them. Figure A1 illustrates the performance and geographical distribution of this subset, comparing the observed storage time series against GWW estimates.

Despite the limited sample size, GWW storage estimates demonstrate relatively high skill, with a median KGE' of 0.65 and 75% of the reservoirs exceeding a score of 0.51. In terms of decomposition, GWW tends to overestimate total storage (median bias of 1.12 with an IQR of 0.13) while slightly underestimating variability (median of 0.91, IQR 0.37). Notably, the temporal dynamics are captured accurately, reflected by a median correlation of 0.88 (IQR 0.10).

Author contributions. JCR designed the study, performed the data processing and analysis, and drafted the initial manuscript. JD assisted with data acquisition and processing and contributed to the manuscript revision. SG assisted in the experimental design and provided critical revisions of the results and the manuscript. PS conceptualized the study and supervised all stages of the research process.

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

Acknowledgements. The authors would like to thank the creators and maintainers of the ResOpsUS, GRanD, and GDW datasets for making their data publicly available. We also acknowledge the use of ERA5 and GloFAS data provided by the European Commission's Copernicus Services.

References

Abeshu, G. W., Tian, F., Wild, T., Zhao, M., Turner, S., Chowdhury, A. F., Vernon, C. R., Hu, H., Zhuang, Y., Hejazi, M., and Li, H. Y.: Enhancing the representation of water management in global hydrological models, Geoscientific Model Development, 16, 5449–5472, https://doi.org/10.5194/gmd-16-5449-2023, 2023.

Baratgin, L., Polcher, J., Dumas, P., and Quirion, P.: Modeling hydropower operations at the scale of a power grid: a demand-based approach, Hydrology and Earth System Sciences, 28, 5479–5509, https://doi.org/10.5194/hess-28-5479-2024, 2024.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W., Heinke, J., Bloh, W. V., and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, Water Resources Research, 47, https://doi.org/10.1029/2009WR008929, 2011.

Burek, P., van der Knijff, J., and de Roo, A.: LISFLOOD. Distributed Water Balance and Flood Simulation Model, Tech. rep., European Commission - Joint Research Centre, Luxembourg, ISBN 879-92-79-33190-9, ISSN 1831-9424, https://doi.org/10.2788/24719, 2013.

Busker, T., De Roo, A., Gelati, E., Schwatke, C., Adamovic, M., Bisselink, B., Pekel, J. F., and Cottam, A.: A global lake and reservoir volume analysis using a surface water dataset and satellite altimetry, Hydrology and Earth System Sciences, 23, 669–690, https://doi.org/10.5194/hess-23-669-2019, 2019.

Choulga, M., Moschini, F., Mazzetti, C., Grimaldi, S., Disperati, J., Beck, H., Salamon, P., and Prudhomme, C.: Technical note: Surface fields for global environmental modelling, Hydrology and Earth System Sciences, 28, 2991–3036, https://doi.org/10.5194/hess-28-2991-2024, 2024.

Coerver, H. M., Rutten, M. M., and Giesen, N. C. V. D.: Deduction of reservoir operating rules for application in global hydrological models, Hydrology and Earth System Sciences, 22, 831–851, https://doi.org/10.5194/hess-22-831-2018, 2018.

Donchyts, G., Winsemius, H., Baart, F., Dahm, R., Schellekens, J., Gorelick, N., Iceland, C., and Schmeier, S.: High-resolution surface water dynamics in Earth's small and medium-sized reservoirs, Scientific Reports, 12, https://doi.org/10.1038/s41598-022-17074-6, 2022.

Duan, Q., Gupta, V. K., and Sorooshian, S.: Shuffled complex evolution approach for effective and efficient global minimization, Journal of optimization theory and applications, 76, 501–521, 1993.

Duan, Q., Sorooshian, S., and Gupta, V. K.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, Journal of Hydrology, 158, 265–284, 1994.

Haddeland, I., Skaugen, T., and Lettenmaier, D. P.: Anthropogenic impacts on continental surface water fluxes, Geophysical Research Letters, 33, https://doi.org/10.1029/2006GL026047, 2006.

Hanasaki, N., Kanae, S., and Oki, T.: A reservoir operation scheme for global river routing models, Journal of Hydrology, 327, 22–41, https://doi.org/10.1016/j.jhydrol.2005.11.011, 2006.

Hanazaki, R., Yamazaki, D., and Yoshimura, K.: Development of a Reservoir Flood Control Scheme for Global Flood Models, Journal of Advances in Modeling Earth Systems, 14, https://doi.org/10.1029/2021MS002944, 2022.

Hao, Z., Chen, F., Jia, X., Cai, X., Yang, C., Du, Y., and Ling, F.: GRDL: A new global reservoir area-storage-depth data set derived through deep learning-based bathymetry reconstruction, Water Resources Research, https://doi.org/10.1029/2023WR035781, 2024.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D.,

Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hou, J., Dijk, A. I. V., Renzullo, L. J., and Larraondo, P. R.: GloLakes: Water storage dynamics for 27000 lakes globally from 1984 to present derived from satellite altimetry and optical imaging, Earth System Science Data, 16, 201–218, https://doi.org/10.5194/essd-16-201-2024, 2024.

Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L.: SPOTting model parameters using a ready-made python package, PLoS ONE, 10, 1–22, https://doi.org/10.1371/journal.pone.0145180, 2015.

JRC: European Flood Awareness System, https://european-flood.emergency.copernicus.eu/en, 2026a.

JRC: Global Flood Awareness System, https://global-flood.emergency.copernicus.eu/, 2026b.

JRC: Open Source Lisflood, https://ec-jrc.github.io/lisflood/, 2026c.

Khandelwal, A., Karpatne, A., Ravirathinam, P., Ghosh, R., Wei, Z., Dugan, H. A., Hanson, P. C., and Kumar, V.: ReaLSAT, a global dataset of reservoir and lake surface area variations, Scientific Data, 9, https://doi.org/10.1038/s41597-022-01449-5, 2022.

Khazaei, B., Read, L. K., Casali, M., Sampson, K. M., and Yates, D. N.: GLOBathy, the global lakes bathymetry dataset, Scientific Data, 9, https://doi.org/10.1038/s41597-022-01132-9, 2022.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Scientific Data, 10, https://doi.org/10.1038/s41597-023-01975-w, 2023.

Kumar, R., Samaniego, L., and Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resources Research, 49, 360–379, https://doi.org/10.1029/2012WR012195, 2013.

Lehner, B., Liermann, C. R., Revenga, C., Vörömsmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., and Wisser, D.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Frontiers in Ecology and the Environment, 9, 494–502, https://doi.org/10.1890/100125, 2011.

Lehner, B., Beames, P., Mulligan, M., Zarfl, C., De Felice, L., van Soesbergen, A., Thieme, M., Garcia de Leaniz, C., Anand, M., Belletti, B., Brauman, K. A., Januchowski-Hartley, S. R., Lyon, K., Mandle, L., Mazany-Wright, N., Messager, M. L., Pavelsky, T., Pekel, J.-F., Wang, J., Wen, Q., Wishart, M., Xing, T., Yang, X., and Higgins, J.: The Global Dam Watch database of river barrier and reservoir information for large-scale applications, Scientific Data, 11, 1069, https://doi.org/10.1038/s41597-024-03752-9, 2024.

Liebe, J., van de Giesen, N., and Andreini, M.: Estimation of small reservoir storage capacities in a semi-arid environment, Physics and Chemistry of the Earth, 30, 448–454, https://doi.org/10.1016/j.pce.2005.06.011, 2005.

Liu, Y., Xie, X., Wang, Y., Tursun, A., Peng, D., Wu, X., and Xue, B.: Increased surface water evaporation loss induced by reservoir development on the Loess Plateau, Hydrology and Earth System Sciences, 30, 67–89, https://doi.org/10.5194/hess-30-67-2026, 2026.

Mekonnen, M. M. and Hoekstra, A. Y.: The blue water footprint of electricity from hydropower, Hydrology and Earth System Sciences, 16, 179–187, https://doi.org/10.5194/hess-16-179-2012, 2012.

Moreno-Rodenas, A., Mantilla-Jones, J. D., and Valero, D.: Age, climate and economic disparities drive the current state of global dam safety, Nature Water, https://doi.org/10.1038/s44221-025-00402-1, 2025.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Pekel, J. F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, 418–422, https://doi.org/10.1038/nature20584, 2016.

Roo, A. P. D., Wesseling, C. G., and Deursen, W. P. V.: Physically based river basin modelling within a GIS: The LISFLOOD model, Hydrological Processes, 14, 1981–1992, https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::aid-hyp49>3.0.co;2-f, 2000.

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

Salamon, P., Grimaldi, S., Mazzetti, C., Prudhomme, C., Russo, C., Zsoter, E., Casado-Rodríguez, J., de Wiart, C. C., Disperati, J., Mastrantonas, N., Azhar, M., Gomes, G., Schweim, C., Sperzel, T., Lemke, C.-D., Ziese, M., Serratosa, A., Jacobson, T., Moschini, F., Bisselink, B., Bavera, D., Ficchì, A., Radke-Fretz, M., and Jiménez-Molina, A.: Improving hydrological modelling and prediction at the European and Global scale, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-8642, https://doi.org/10.5194/egusphere-egu25-8642, 2025.

Salwey, S., Coxon, G., Pianosi, F., Lane, R., Hutton, C., Bliss Singer, M., McMillan, H., and Freer, J.: Developing water supply reservoir operating rules for large-scale hydrological modelling, Hydrology and Earth System Sciences, 28, 4203–4218, https://doi.org/10.5194/hess-28-4203-2024, 2024.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 1–25, https://doi.org/10.1029/2008WR007327, 2010.

Scherer, L. and Pfister, S.: Global water footprint assessment of hydropower, Renewable Energy, 99, 711–720, https://doi.org/10.1016/j.renene.2016.07.021, 2016.

Schwatke, C., Dettmering, D., Bosch, W., and Seitz, F.: DAHITI - An innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry, Hydrology and Earth System Sciences, 19, 4345–4364, https://doi.org/10.5194/hess-19-4345-2015, 2015.

Schwatke, C., Dettmering, D., and Seitz, F.: Volume variations of small inland water bodies from a combination of satellite altimetry and optical imagery, Remote Sensing, 12, https://doi.org/10.3390/rs12101606, 2020.

Shen, H., Tolson, B. A., and Mai, J.: Time to Update the Split-Sample Approach in Hydrological Model Calibration, Water Resources Research, 58, https://doi.org/10.1029/2021WR031523, 2022.

Shen, Y., Yamazaki, D., Pokhrel, Y., and Zhao, G.: Improving Global Reservoir Parameterizations by Incorporating Flood Storage Capacity Data and Satellite Observations, Water Resources Research, https://doi.org/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 Resources Research, 55, 787–810, https://doi.org/10.1029/2018WR023025, 2019.

Shrestha, P. K., Samaniego, L., Rakovec, O., Kumar, R., Mi, C., Rinke, K., and Thober, S.: Toward Improved Simulations of Disruptive Reservoirs in Global Hydrological Modeling, Water Resources Research, 60, https://doi.org/10.1029/2023WR035433, 2024.

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, Hydrology and Earth System Sciences, 28, 1071–1088, https://doi.org/10.5194/hess-28-1071-2024, 2024.

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

Steyaert, J. C., Sutanudjaja, E. H., Bierkens, M., and Wanders, N.: Data derived reservoir operations simulated in a global hydrologic model, Hydrology and Earth System Sciences, 29, 6499–6527, https://doi.org/10.5194/hess-29-6499-2025, 2025.

Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., and Samaniego, L.: The multiscale routing model mRM v1.0: Simple river routing at resolutions from 1 to 50 km, Geoscientific Model Development, 12, 2501–2521, https://doi.org/10.5194/GMD-12-2501-2019, 2019.

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

US Army Corps of Engineers: National Inventory of Dams, https://nid.sec.usace.army.mil/{#}/, 2025.

US Bureau of Reclamation: Projects & Facilities, https://www.usbr.gov/projects/facilities.php?type=Dam, 2025.

US Geological Survey: Water Data for the Nation, https://waterdata.usgs.gov/, 2025.

van der Knijff, J. M., Younis, J., and de Roo, A. P.: LISFLOOD: A GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, https://doi.org/10.1080/13658810802549154, 2010.

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, Hydrology and Earth System Sciences, 23, 3735–3764, https://doi.org/10.5194/hess-23-3735-2019, 2019.

Yigzaw, W., Li, H. Y., Demissie, Y., Hejazi, M. I., Leung, L. R., Voisin, N., and Payn, R.: A New Global Storage-Area-Depth Data Set for Modeling Reservoirs in Land Surface and Earth System Models, Water Resources Research, 54, 10,372–10,386, https://doi.org/10.1029/2017WR022040, 2018.

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

Zhao, G. and Gao, H.: Automatic Correction of Contaminated Images for Assessment of Reservoir Surface Area Dynamics, Geophysical Research Letters, pp. 6092–6099, https://doi.org/10.1038/nature20584, 2018.