https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

Using satellite observations to validate and improve reservoir storage simulations in global hydrological models

Emmanuel Okiria[1], Naota Hanasaki[1,2], Simon N. Gosling[3], Emmanuel Nyenah[4,5], Peter Burek[6], Yusuke Satoh[7], Sebastian Ostberg[8], Kedar Otta[1], Luca Guillaumot[6,9]

[1]National Institute for Environmental Studies (NIES), Tsukuba, 305-8506, Japan
[2]School of Engineering, The University of Tokyo, Tokyo, 113-8656, Japan
[3]School of Geography, University of Nottingham, Nottingham, NG7 2RD, United Kingdom
[4]Institute of Physical Geography, Goethe University, Frankfurt, 60438 Frankfurt am Main, Germany
[5]Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, 60325 Frankfurt am Main, Germany
[6]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[7]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, 237-0061, Japan
[8]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
[9]BRGM - French Geological Survey, F-45060 Orléans, France

Correspondence to: Emmanuel Okiria (okiria.emmanuel@nies.go.jp)

Abstract. Global hydrological models (GHMs) increasingly incorporate *generic reservoir operation schemes* (GROS) to simulate the regulation of rivers by dams. However, the reliability of GROS remains largely unvalidated on a global scale due to the historical scarcity of open *in situ* data. Here, we leverage the Global Reservoir Storage (GRS) satellite dataset to conduct the first comprehensive quantitative evaluation of reservoir storage simulations globally from five GHMs: H08, WaterGAP2-2e (WGP), MIROC-INTEG-LAND (MIL), CWatM (CWT) and LPJmL5-7-10-fire (LPJ). H08, WGP, MIL and LPJ adopted the process-based Hanasaki *et al*. (2006) reservoir operation scheme (H06), while CWT adopted the piecewise-function rule curve approach of Burek *et al*. (2013, 2020) (LIS). We address two primary questions: (1) how accurately do state-of-the-art GHMs reproduce global reservoir storage dynamics? and (2) are model deficiencies attributable to parametric rigidity (*i.e*., the adoption of globally uniform parameters) in GROS? We evaluated monthly reservoir storage series at 424 major dams (capacity ≥ 0.5 km³) over the historical period, 1999–2018. Performance was quantified using the Kling-Gupta Efficiency (KGE). Two *post-hoc* bias correction methods—linear scaling and variance-matching—were applied to the raw monthly storage simulations to evaluate whether simple, targeted statistical transformations could recover model skill. To comprehensively address parametric rigidity, we conducted a sensitivity analysis on H08 using its H06 scheme by varying two parameters: *target storage level* (TSL) and the *degree of regulation threshold* (DORT) and using LIS by varying the *normal storage limit* (LN). Our evaluation reveals that current GROS yield generally unsatisfactory performance, characterised by two distinct features. The first concerns seasonal amplitude in storage. MIL initially achieves the highest skill: 52.36% of dams had a KGE > -0.41. However, KGE decomposition revealed this skill was largely due to dampened intra-annual variability rather than being driven by high correlation and/or low bias error. In contrast, the other GHMs often exhibit excessive seasonal drawdown, systematically overestimating storage amplitude. The second feature pertains to temporal dynamics in storage: within the group exhibiting exaggerated seasonal drawdown, H06-based models—H08, WGP and LPJ—significantly outperform the LIS-based CWT in temporal correlation. We demonstrate that when variance-matching bias correction is applied across all GHMs, two things happen: firstly, the performance of all GHMs becomes generally satisfactory (median KGE > -0.41), and secondly, the GHMs with exaggerated seasonal drawdown outperform MIL in terms of KGE, owing to their superior temporal correlation (H06-based GHMs) and mean bias estimation performance (except H08). By contrast, linear scaling yields only marginal improvements, indicating that correcting variability errors is substantially more effective than adjusting mean bias alone. Furthermore, sensitivity analyses confirm that exaggerated seasonal drawdown is primarily a result of parameter choices rather than inherent flaws in GROS. These findings highlight two critical insights: (1) one-size-fits-all parameters are a primary limitation in global reservoir modelling; and (2) satellite observations are a viable dataset for calibrating reservoir operation schemes in GHMs.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

1. Introduction

Dams constitute one of the most consequential anthropogenic interventions in the terrestrial water cycle (Poff *et al*. 1997; Nilsson *et al*. 2005; Grill *et al*. 2019). By redistributing natural discharge variability to satisfy irrigation, industrial, and domestic water demands, as well as hydropower generation and flood control, they fundamentally alter downstream hydrological regimes and ecosystem connectivity (Poff *et a*l. 1997; Nilsson *et al*. 2005). To quantify and /or represent these impacts on a planetary scale, *global hydrological models* (GHMs) incorporate reservoir operation schemes. In most GHMs, reservoir operations are governed by *generic reservoir operation schemes* (GROS), such as the methods by Hanasaki *et al*. (2006) (hereafter H06), Döll *et al*. (2003) (hereafter D03) and Burek *et al*. (2013; 2020) (hereafter LIS), among others. However, the reliability of these simulations (from GROS) remains a critical area of uncertainty at the global scale (Gao *et al*. 2012; Nazemi and Wheater 2015a; 2015b; Hosseini-Moghari and Döll 2025). Moreover, while GHMs are routinely calibrated and validated against river discharge, the validation of reservoir storage—one of the state variables—has been historically neglected due to the proprietary nature of *in situ* data (Gao *et al*. 2012; Zajac *et al*. 2017; Busker *et al*. 2019). Consequently, while a GHM may successfully reproduce downstream flows through calibration against observed discharge, it may still fail to accurately capture the storage dynamics. Such storage-discharge parameter optimisation trade-offs were reported in Yassin *et al*. (2019); Döll *et al*. (2024); Hasan *et al*. (2025); and Hosseini-Moghari and Döll (2025). This discrepancy undermines the physical consistency of the simulations and limits the reliability of subsequent climate-impact assessments.

To address the uncertainty in storage simulations by GROS, studies have adopted two distinct pathways: deriving reservoir operating curves from observed data and performing diagnosis, and refinement of the existing GROS. The first pathway (data-derived operating curves) is an alternative to GROS. Early efforts were confined to data-rich areas, most notably, the conterminous United States (CONUS). Using the ResOpsUS dataset (Steyaert *et al*. 2022), Turner *et al*. (2021) inferred upper and lower storge targets for individual reservoirs in the CONUS, effectively reconstructing operating curves from observed behaviour. Building on this approach, Steyaert *et al*. (2025) extended data-derived operating curves to a global scale by leveraging satellite-derived storage estimates from GloLakes (Hou *et al*. 2024). For the ~22,000 reservoirs in the GeoDAR dataset (Wang *et al*. 2022) lacking direct observations, they inferred the operating bounds using a *random forest* model trained on the 1,752 observed dams. Steyaert *et al*. (2025) subsequently implemented these data-derived operating curves in PCR-GLOBWB 2 (Sutanudjaja *et al*. 2018), yielding improved performance relative to the GHM's original GROS.

However, despite the promise of data-driven methods, their application has so far been restricted to regional studies relying on *in situ* observations. To our knowledge, the only global application to date has been limited to a single model framework, PCR-GLOBWB 2 (Steyaert *et al.* 2025). Consequently, GROS continue to underpin most GHMs used in global water resources and climate impact assessments. This reality motivates the second research pathway: the systematic diagnosis of the existing GROS, rather than their replacement. Efforts to validate GROS have been ongoing, albeit limited in scope. Masaki *et al*. (2017) pioneered the multi-GHM intercomparison of GROS for five GHMs but their assessment was geographically restricted to two river basins in the CONUS and relied on validating discharge downstream of dams to infer the performance of GROS indirectly, rather than validating the storage dynamics directly. Later, Gutenson *et al*. (2020) compared two GROS (H06 and D03), expanding coverage to 60 dams in the CONUS. But again, their study focused on evaluating the accuracy of dam-influenced discharge simulations regionally. In a move towards evaluation of both storage and release by GROS, Sadki *et al*. (2023) evaluated the combined storage-release performance of a modified (a more heavily parameterised) H06—so-called Dam-Reservoir Operation (DROP) (Sadki *et al*. 2023)—against *in situ* data across 215 reservoirs in Spain. Critically, it was a direct evaluation of the GROS, since the inflows to DROP were reconstructed from observed data. Driven by the growing accessibility of satellite-derived datasets, recent studies, *e.g.*, Otta *et al*. (2025), and Hosseini-Moghari and Döll (2025) have attempted to directly validate storage simulations against satellite-derived storage benchmarks. However, these studies face distinct limitations regarding their quantitative rigor and/ or scope. Otta *et al*. (2025) relied on normalised storage time series to overcome data uncertainties, rendering their assessment largely qualitative—focusing on temporal correlation—rather than

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

a quantitative absolute storage evaluation. Moreover, they compared only two GHMs—H08 (Hanasaki *et al*. 2018) and WaterGAP2-2e (Müller Schmied *et al*. 2024)—and their pilot study was limited to seven dams in the CONUS. While Hosseini-Moghari and Döll (2025) performed a quantitative evaluation (evaluated absolute volumetric time series), they were constrained to 100 CONUS reservoirs and analysed only a single GHM, WaterGAP2-2e.

Collectively, while earlier studies demonstrate the feasibility and value of absolute storage-based evaluation, they reveal a persistent gap: no study has performed a global, quantitative assessment of reservoir storage simulations across multiple GHMs. Consequently, it remains unclear how the standard GROS—which still underpin most climate impact assessments—perform globally when confronted with absolute volumetric observations. To address this gap, this study presents the first global, quantitative evaluation of reservoir storage simulations from several GHMs against a comprehensive satellite-derived dataset, the Global Reservoir Storage (Li *et al*. 2023). The study considers five GHMs from the Inter-Sectoral Impact Model Intercomparison Project Phase 3a (ISIMIP3a) (Frieler *et al*. 2024): H08, WaterGAP2-2e, MIROC-INTEG-LAND (Yokohata *et al*. 2020), CWatM (Burek *et al*. 2020) and LPJmL5-7-10-fire (Wirth *et al*. 2024; Oberhagemann *et al*. 2025). Specifically, we seek answers to two key questions: (1) how accurately do these state-of-the-art GHMs reproduce global reservoir storage time series when benchmarked against satellite-derived storage observations? and (2) to what extent are GHM deficiencies attributable to parametric rigidity in generic schemes, and can these limitations be alleviated through targeted statistical bias correction or parameter tuning? By answering these questions, we aim to provide the diagnostic foundation necessary for the improvement of storage simulations by GHMs.

Note that throughout this paper, the terms “dam” and “reservoir” are used interchangeably, as each dam impounds its corresponding reservoir, *i.e*., only reservoirs that are associated with a dam will be studied.

2. Materials and methods

2.1. The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP)

This study utilises simulations from the Inter-Sectoral Impact Model Intercomparison Project Phase 3a (ISIMIP3a) (Frieler *et al*. 2024), a standardised framework designed to ensure consistent cross-sectoral global-scale climate change impact assessments. Within the Global Water sector of ISIMIP3a, participating GHMs were driven by harmonised, bias-adjusted climate forcings described by Lange *et al*. (2022). To account for climate forcing uncertainty, we considered simulation runs that used three distinct datasets of observed climate as model input—GSWP3-W5E5, 20CRv3-W5E5 and 20CRv3-ERA5—provided at a daily temporal resolution on a 30-arcminute global grid. While the sector (Global Water sector) includes 15 participating GHMs, this study evaluates the subset of five—H08 (Hanasaki *et al*. 2018), WaterGAP2-2e (WGP) (Müller Schmied *et al*. 2024), MIROC-INTEG-LAND (MIL) (Yokohata *et al*. 2020), CWatM (CWT) (Burek et al., 2020) and LPJmL5-7-10-fire (LPJ) (Wirth *et al*. 2024; Oberhagemann *et al*. 2025)—that explicitly simulate river regulation by dams and provided the storage outputs at the time this study was conducted. For each of the reservoirs analysed in this study (capacity ≥ 0.5 km³: see Fig. 1), 15 distinct storage time series were obtained (three climate forcings and five GHMs). For each GHM, we then averaged simulations across the three forcings to produce one representative time series per reservoir, per GHM. To account for forcing-induced uncertainty, we present individual simulation results in the Supplementary, allowing us to quantify the spread and validate the robustness of the conclusions. Note that for H08, we utilise a corrected simulation run performed for this study specifically, rather than the simulation available on the ISIMIP3a repository (see Sect. 2.3.3 for details). All storage simulations are monthly series, covering the historical period (spanning 1901–2019/2021) determined by the temporal coverage of the respective climate forcings. To ensure topological consistency, all GHMs adopted the standard ISIMIP3a reservoir input dataset. This dataset harmonises the locations and capacities, among other static properties, of 7,330 dams—7,319 dams from the GRanD v1.3 inventory (Lehner *et al.* 2011) and 11 dams provided by Dr. Jida Wang, Kansas State University (KSU)—ensuring that all GHMs simulate the same dams. The dam locations were mapped onto the DDM30 river

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

network of Döll and Lehner (2002), a global drainage direction map at 30-arcminute resolution for application in GHMs. A more detailed description of ISIMIP3a can be found at https://protocol.isimip.org and Frieler *et al*. (2024).

2.2. Global hydrological models and reservoir operation schemes

The multi-model ensemble comprises five *global hydrological models* (GHMs): H08, MIL, WGP, LPJ and CWT. While these models share basic hydrological principles, they exhibit significant structural diversity and conceptual differences (Telteu *et al*. 2021; Müller Schmied *et al*. 2025), as summarised in Table 1.

Table 1: Summary of the reservoir schemes, reservoir classifications and parameter tuning status across the ISIMIP GHMs used in thus study.

GHM	Reservoir Scheme	Runoff Routing Scheme	Global/Local Reservoir Classification	Land Surface Hydrology Module Calibration (against observed discharge)
H08	H06 (Hanasaki *et al*. 2006). (H06 classifies reservoirs by primary purpose as irrigation or non-irrigation, and applies distinct release algorithms for each class)	Linear flow routing (constant velocity)	Global: Catchment area > 5,000 km^2 Local: Catchment area ≤ 5,000 km^2	Calibrated (regionalised by climate zone)
MIL	H06	Linear flow routing (constant velocity)	Global: Capacity ≥ 1 km^3 Local: Capacity < 1 km^3	Uncalibrated
WGP	H06	Linear flow routing (variable velocity)	Global: Capacity ≥ 0.5 km^3 Local: Capacity < 0.5 km^3	Calibrated
LPJ	H06	Linear flow routing (constant velocity)	No global/local separation	Uncalibrated
CWT	LISFLOOD reservoir scheme (LIS) (Burek *et al*., 2013; 2020) (Reservoirs are not classified by purpose and are simulated using the same release algorithm)	Kinematic-wave routing	No global/local separation	Uncalibrated

A substantial commonality across the ensemble is the reliance on the H06 (Hanasaki *et al*. 2006) reservoir operation algorithm: four of the five models employ the H06 logic—or its variants—to simulate the regulation of *global* reservoirs. Conceptually, *global* reservoirs are those located on the main stem of major rivers that are explicitly represented on the global digital river network used by the GHMs (DDM30 for ISIMIP3a GHMs), whereas *local* reservoirs are situated on tributaries of the main rivers (Hanasaki *et al*. 2018). The quantitative description of global and local reservoirs can be found in Table 1. A fundamental description of the H06 release formulation is shown in Sect. 2.4.4 while a complete description can be found in Hanasaki *et al*. (2006). Furthermore, they (the four H06-based GHMS) adopt a linear flow routing scheme, with H08 and MIL adopting the constant flow velocity Total Runoff Integrating Pathways (TRIP) (Oki and Sud, 1998; Oki *et al*. 1999) method and WGP adopting a variable velocity from the Manning-Strickler approach. Despite these similarities, these four models diverge significantly in their classification of reservoirs (as *global* or *local* reservoirs) and parameterisation strategies: note that H06 or LIS are applied to only *global* reservoirs. H08 defines *global* reservoirs based on a catchment area threshold of 5,000 km²,

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

treating the rest (*local* reservoirs) as ideal tanks. In contrast, MIL and WGP utilise storage capacity thresholds of 1 km³ and 0.5 km³, respectively. Both MIL and WGP simulate *local* reservoirs as natural lakes, with WGP specifically applying the hydraulic weir relationship of Döll *et al*. (2003) to these reservoirs (*local* reservoirs) (Müller Schmied *et al*. 2021).

LPJ adopts a hybrid reservoir operation strategy without a *global-local* reservoir classification (Biemans *et al*. 2011). Consequently, all reservoirs are simulated using the same reservoir operation scheme. To define reservoir release, it (LPJ) retains the H06 framework, with its default global parameter values as defined in Hanasaki *et al*. (2006). Spatially, it adopts the logic of Biemans *et al*. (2011), extending the command area to include grid cells within a 5-cell (~250 km at the equator) radius upstream of the main river to simulate water conveyance. Crucially, LPJ optimises releases from irrigation reservoirs exclusively to satisfy the aggregated irrigation deficits of this command area, while domestic and industrial water demands can only be satisfied using locally available water resources.

The ensemble includes one structural outlier regarding reservoir operations and runoff routing, CWT. While the other GHMs adopt H06, CWT implements the LIS scheme (Burek *et al*. 2013; 2020). In this scheme, releases follow a piecewise-function-defined operating rule based on three storage level thresholds: *conservative*, *normal* and *flood limits* storage level thresholds (Eq. (S3)). Moreover, it is also unique in its use of the kinematic-wave approximation of the Saint-Vanant equation (Chow *et al*. 1988) for routing runoff.

Regarding parameter tuning, WGP and H08 are the only GHMs in this ensemble that apply calibration to their land surface parameters against observed river discharge. Specifically, WGP adopts basin-scale calibration at observed discharge station locations, whereas H08 regionalises its parameters by climate zone, using the method proposed by Yoshida *et al*. (2022). The remaining GHMs have land surface processes that are not calibrated against observed discharge.

2.3. Data acquisition and processing

2.3.1 Satellite-based storage observations

As mentioned earlier, with the limited global coverage of ground-based networks and the paucity of openly available ground observations, satellite-derived storage observations have emerged as potential global surrogates for *in situ* data. This section details the methodology used to intercompare three satellite-derived storage datasets against *in situ* measurements. The primary goal is to evaluate their appropriateness as surrogates and to identify the best performing satellite product to serve as a benchmark for evaluating simulated storage from GHMs.

Reservoir storage observations were obtained from two sources: *in situ* measurements and satellite-derived products. Monthly *in situ* storage time series, considered the most reliable ground truth, were compiled across five countries (Australia, Canada, India, Spain and the United States), yielding a total of 1,081 reference dams, all of which are included in the GRanD v1.3 database. However, due to the limited global coverage of ground-based networks, and the paucity of openly available ground observations, we explored the potential of satellite-derived products as global surrogates for *in situ* data. We collected three latest satellite-derived reservoir storage products—Global Reservoir Storage (GRS) (Li *et al.* 2023), GloLakes (Hou *et al.* 2024) and Global Dam Storage (GDS) (a new dataset developed in this study, see Sect. 2.3.2)—and validated them against *in situ* monthly storage series. From GRS, GloLakes and GDS, we obtained storage time series data from 7,245, 3,890 and 7,229 dams respectively (all having GRanD IDs). Of these, 774 reservoirs intersected across all three satellite products and the 1,081 *in situ* reference dams. To ensure robust statistical comparisons, we required a minimum of 60 months (5 years) of overlapping data between the *in situ* and satellite-derived records, which narrowed the pool to 638 dams (blue and green markers in Fig. S1 of the Supplementary). These were used for the direct satellite product inter-comparison with *in situ* data.

To illustrate the performance of the satellite-derived products, four specific dams—green markers in Fig. S1—were selected for detailed time series analysis: Libby (the Americas), Ebro (Europe), Rana Pratap Sagar (Asia) and Wyangala (Oceania). The selection of these specific reservoirs was based on three criteria: (1) the availability of continuous *in situ* storage time series, overlapping for at least 60 months with all three satellite products; (2) geographic diversity, representing each major

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

United Nations geoscheme macro region (United Nations Statistics Division, n.d.) with *in situ* data; and (3) statistical representativeness, whereby the performance metrics of selected reservoirs closely match the median behaviour of the full validation dataset, as reflected in the performance metrics. This selection ensures that the analysed time series provide a realistic and unbiased illustration of the overall performance characteristics of the satellite products.

2.3.2 Development of the Global Dam Storage (GDS) satellite-storage dataset

In an attempt to explore potential improvements in satellite-derived storage estimations using the recently published GRDL (Hao *et al*. 2024) bathymetry dataset, we developed a new satellite-based dataset: Global Dam Storage (GDS). GDS was developed by integrating area time series from the Global Reservoir Surface Area Dataset (GRSAD) (Zhao and Gao, 2018) with GRDL bathymetry. To constrain bathymetric curves derived from GRDL, we established reference capacity values for each reservoir—volume, area and depth at capacity—from the Global Dam Watch (GDW) dataset Lehner *et al.* (2024). When GDW reported a larger storage capacity than GRDL, the area-depth and depth-volume pairs were augmented accordingly to ensure that the hypsometric curves spanned the entire bathymetry of the reservoir. Storage was then estimated by fitting multiple candidate models—piecewise linear interpolation, cubic splines and polynomial functions—to the combined area–volume and area–depth data. Following the geometry optimisation framework proposed by Yigzaw *et al*. (2018) and later adapted by Li *et al*. (2023), we selected the optimal bathymetric model for each reservoir by identifying the one that minimised the combined relative error between our estimated values and the GDW-reported storage and depth at capacity. Further details regarding the derivation of GDS, specifically the selection of the optimal reservoir bathymetry are provided in Sect. S1 of the Supplementary. For completeness, an evaluation of GHM storage simulations against GDS is provided in Sect. S8.

2.3.3 Simulated reservoir storage data

The simulated reservoir storage data were obtained from the five ISMIP3a GHMs. We screened the simulated reservoir storage as follows. First, the analysis included only artificial reservoirs (dam-created reservoirs), excluding partially regulated lakes, *e.g*. Lake Victoria in East Africa. Furthermore, reservoirs whose DDM30-derived catchment area deviated by more than 30% from the *true* observed catchment area reported in GDW were excluded from further analysis. We then identified reservoirs present in both the GRS dataset and all five ISIMIP3a GHMs, yielding an initial set of 484 dams. Finally, we excluded any reservoirs for which the Kling-Gupta Efficiency (Sect. 2.4.1) was undefined in any of the models, resulting in a core sample of 424 dams (Fig. 1). Note that GRS is chosen because it performs best when compared to *in situ* data (Sect. 3.1).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

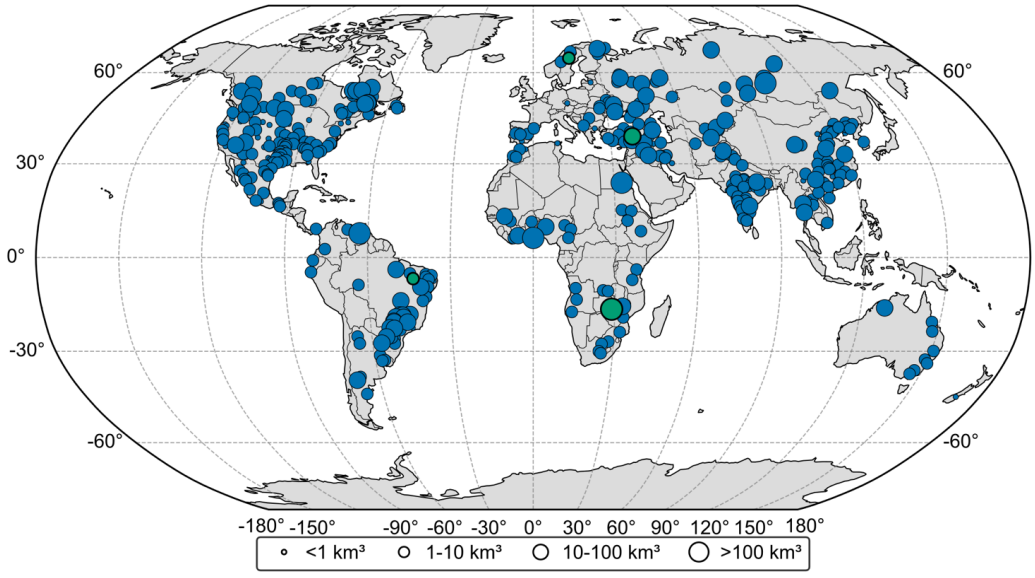


Figure 1: Global distribution of the 424 dams included in this study, for which reservoir storage is simulated by the ISIMIP3a GHMs. Marker size is proportional to reservoir storage capacity. Green markers indicate the dams selected for detailed time series analysis: Boa Esperança (the Americas), Umluspen (Europe), Kariba (Africa) and Keban (Asia).

We also created a new model configuration—termed H08-LIS—by integrating the LIS reservoir scheme into H08. Running the H08-LIS configuration following the ISIMIP3a protocol allowed us to infer whether performance differences between H06-based models and the LIS-based CWT were driven by the reservoir schemes themselves or by the host-model inflows into the reservoir.

Note that during preliminary diagnostics, we found and fixed a minor bug in the reservoir inflow calculation in H08. Consequently, the reservoir variables that we analyse in this paper and those in the ISIMIP repository are different for reservoirs with a *degree of regulation* (DOR) less than 0.5. Quantitative evaluation/impacts of this correction are reported in Sect. S4 of the Supplementary (Fig. S3–S6). Table 2 summarises the primary datasets used in this study.

Table 2: A summary of the datasets used in this study.

Dataset	Data Type	Main Application in this Study	Reference(s)
Global Reservoir and Dam database (GRanD)	Dam and reservoir metadata	Provides the database for dams evaluated in ISIMIP3a GHMs	Lehner *et al.* (2011)
Global Dam Watch (GDW)	Dam and reservoir metadata	Provides reference reservoir values at capacity (volume, area and depth) used to derive GDS, and true catchment areas used to screen simulated reservoirs	Lehner *et al.* (2024)
Global Reservoir Surface Area Dataset (GRSAD)	Satellite-derived reservoir surface area time series	Integrated with GRDL bathymetry to develop the new GDS dataset	Zhao and Gao, (2018)
GRDL	Reservoir bathymetry dataset	Integrated with GRSAD to develop the new GDS dataset: provides bathymetric curves constrained by reference capacity values to derive volumetric storage	(Hao *et al.* 2024)

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

Global Reservoir Storage (GRS)	Satellite-derived reservoir storage time series	Evaluated as a candidate surrogate for *in situ* storage data	Li *et al*. (2023)
GloLakes	Satellite-derived reservoir storage time series	Evaluated as a candidate surrogate for *in situ* storage data	Hou *et al.* (2024)
Global Dam Storage (GDS)	Satellite-derived reservoir storage time series	Evaluated as a candidate surrogate for *in situ* storage data	This Study

Similarly, to illustrate the performance of the five ISIMIP3a GHMs against the satellite-storage benchmark, four representative dams were selected for detailed time series analysis (green markers in Fig. 1): Boa Esperança (the Americas), Umluspen (Europe), Kariba (Africa) and Keban (Asia). The selection was based on three criteria: (1) data availability and model overlap, whereby each reservoir was required to have continuous satellite-derived storage time series, be simulated by all 5 ISIMIP3a GHMs, and provide at least 60 months of overlapping satellite-GHM storage data; (2) geographic diversity, ensuring representation across at least four of the five major global regions (United Nations Statistics Division, n.d.); and (3) statistical representativeness, such that the performance metrics of the selected reservoirs closely match the median behaviour of the ISIMIP3a GHMs across the global dataset (N = 424 dams). This selection ensures that the analysed storage time series provide realistic and unbiased illustration of typical ISIMIP3a GHM performance.

2.4. Model evaluation and improvement

To systematically evaluate and improve the reservoir storage simulations, our analytical methodology was structured into four steps. First, we conducted a multi-metric evaluation to benchmark model performance against observed data (Step 1). Second, we applied an additive time-series decomposition to analyse temporal storage dynamics and isolate seasonal cycles (Step 2). Third, we implemented targeted *post-hoc* bias correction to determine if simple statistical bias correction could recover model skill (Step 3). Finally, we performed sensitivity analyses and parameter tuning on the models' reservoir operation schemes to address underlying parametric errors (Step 4).

2.4.1 Evaluation metrics

To assess the fidelity of satellite-derived storage time series against *in situ* data as well as GHM reservoir storage simulations against the satellite-derived storage benchmark, we implemented a multi-metric evaluation framework. The primary metric was the Kling-Gupta Efficiency (KGE) (Eq. (1)), selected for its ability to jointly evaluate correlation, variability and bias errors. We adopted KGE = -0.41 as the threshold for a skilful model, following the works of Knoben *et al*. (2019). KGE was decomposed into its three components—the Pearson correlation coefficient (R), the variability ratio (α) and the bias ratio (β)—to attribute model errors to timing, amplitude, or mean bias.

$$KGE = 1 - \sqrt{(R-1)^2 + (\alpha-1)^2 + (\beta-1)^2}\,, \quad (1)$$

$$\alpha = \frac{\sigma_{sim}}{\sigma_{obs}}\,, \quad (2)$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}}\,, \quad (3)$$

where, μ_{obs} and μ_{sim} denote observed and simulated means respectively, while σ_{obs} and σ_{sim} represent their respective standard deviations.

To complement KGE, we used the Symmetric Mean Absolute Percentage Error (SMAPE) (Flores 1986; Makridakis *et al*. 1993) (Eq. (4)) to quantify relative error. Unlike MAPE, which is asymmetrical (for over- vs. under-prediction) and becomes unstable when observed values are close to zero, SMAPE is more stable because it normalises errors by the average of the

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

observed and simulated values rather than the observed values alone. As a result, when observed reservoir storage approaches zero, SMAPE does not erratically result in extremely large values, making it a robust metric, even for low storage conditions.

$$SMAPE = \frac{2\times100}{n}\sum_{i=1}^{n}\frac{|obs_i - sim_i|}{|obs_i| + |sim_i|} \quad (4)$$

2.4.2 Time series decomposition

To analyse the temporal dynamics of reservoir water storage, we employed an additive decomposition method. The monthly storage series is decomposed into three distinct components: the long-term annual mean, mean annual seasonal cycle and residuals (Eq. (5)). Decomposition is essential because it isolates the seasonal response of time series within a calendar year, allowing for a clear comparison of how well GHMs reproduce the seasonal cycle and the inter-annual variability.

$$S_{y,m} = S_y + \bar{S}_m + e_{y,m} \quad (5)$$

Where $S_{y,m}$ is the monthly storage time series for month m in year y, S_y is the annual average storage for year y and $\bar{S}_m$ is the seasonal variability. $\bar{S}_m$ is determined by first subtracting S_y from $S_{y,m}$, and then calculating the long-term mean of these deviations for each calendar month, *i.e.*, averaging all Januarys, all Februarys, *etc*. Finally, $e_{y,m}$ are residuals. These are the component which cannot be attributed to either annual average storage or seasonal variability. Prior to decomposition, the time series data were filtered to remove the filling period of the reservoirs. Using the dam construction year provided in the GDW dataset, we masked all storage data prior to the construction year, the construction year itself, and the subsequent five calendar years. This ensures that the decomposition analysis is performed strictly on the stabilised operational phase of the reservoirs, avoiding artifacts from the initial impoundment period.

2.4.3 Bias correction of storage time series

To systematically investigate the sources of model error, we applied two *post-hoc* bias correction methods to the original monthly storage simulations. This intermediate step was designed as an artificial diagnostic test. By isolating and correcting mean bias and variability independently, we aimed to determine exactly how much these specific components degrade overall model skill before attempting to fix them. The first method, linear bias correction (Eq. (6)), addressed systematic storage errors by scaling the simulated storage to match the observed storage mean. The second method, variance-matching (Eq. (7)), corrects the misrepresented storage variability, while intentionally preserving the simulated storage mean.

$$x' = \frac{\mu_{obs}}{\mu_{sim}}(x) \quad (6)$$

$$x' = \frac{\sigma_{obs}}{\sigma_{sim}}(x - \bar{x}) + \bar{x} \quad (7)$$

Here, x', x and $\bar{x}$ denote the bias corrected, original and the mean storage time series respectively. Note that applying these transformations affected the number of reservoirs with computable metrics: after variance-matching bias correction, 401 dams had a defined KGE, whereas after linear bias correction, 431 dams retained a defined KGE.

2.4.4 Sensitivity test/parameter tuning

Building on the diagnostic baseline established by the statistical bias corrections, our next planned step was to investigate a concrete, mechanistic solution. We designed sensitivity tests to determine whether parameter tuning could naturally cure the targeted errors (and potentially improve temporal correlation) from within the model itself. To do this, we evaluated the H08 model using its default reservoir operation scheme (H08-H06) and the newly integrated LIS scheme (H08-LIS). Following the ISIMIP3a protocol (https://protocol.isimip.org/#/ISIMIP3a), H08-H06 and H08-LIS were forced with GSWP3-W5E5 climate data for the period 1980-2000. For H08-H06, two key parameters governing reservoir water release decisions were examined:

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

- *Target storage level* (TSL). This appears in Eq. (8), and it controls the mean bias in the release time series, consequently affecting the mean bias of the storage time series. Note that TSL corresponds to Hanasaki *et al*. (2006)'s α parameter: we adopted TSL to avoid confusion with KGE's variability term (α), and because it seemed more self-descriptive.
- *Degree of regulation threshold* (DORT). The *degree of regulation* (DOR) is defined in Eq. (9) as the ratio of the reservoir storage capacity (denoted by upper case C) to long-term mean annual inflow into the reservoir. A high DOR indicates a substantial carry-over capacity: the reservoir can store a large fraction of the annual inflow, and so be able to buffer seasonal extremes, and exert strong, control over releases (we shall call this the *heavily regulated* mode). Conversely, a low DOR indicates limited storage capacity relative to reservoir inflow. Here, the reservoir lacks the buffer to hold water for extended periods, operating more like a run-of-the-river system, where releases track inflow. Its threshold—the *degree of regulation threshold* (DORT)—determines the *degree of regulation* (DOR) at which the reservoir release transitions along the continuum between heavily regulated (Eq. (10)) and inflow-tracking behaviour (Eq. (11)). DORT corresponds to Hanasaki *et al*. (2006)'s c variable (lower case c).

$$\mathrm{Krls} = \frac{S_{1st,y}}{\mathrm{TSL} \times \mathrm{C}} \tag{8}$$

$$\mathrm{DOR} = \frac{\mathrm{C}}{\mathrm{inf_{ann}}} \tag{9}$$

$$\mathrm{Rls} = \mathrm{Krls} \times \mathrm{rls_{fxd}} \tag{10}$$

$$\mathrm{Rls} = \left(\frac{\mathrm{DOR_{sim}}}{\mathrm{DORT}}\right)^2 \times \mathrm{Krls} \times \mathrm{rls_{fxd}} + \left\{1 - \left(\frac{\mathrm{DOR_{sim}}}{\mathrm{DORT}}\right)^2\right\} \times \mathrm{inf_m} \tag{11}$$

In Eq. (8), Krls is the storage adjustment factor (updated at the beginning every reservoir operation year), $S_{1st,y}$ is the storage at the start of the y^{th} reservoir operation year and C (upper case C) is the reservoir storage at capacity. The denominator in Eq. (8), $\mathrm{TSL} \times \mathrm{C}$, represents the desirable or targeted storage volume of the reservoir. Further, $\mathrm{inf_{ann}}$ is the long term mean annual inflow into the reservoir, $\mathrm{rls_{fxd}}$ is the fixed release, which is equivalent to the long-term mean annual inflow into the reservoir (note that $\mathrm{rls_{fxd}}$ and $\mathrm{inf_{ann}}$ are basically the same thing), Rls is reservoir release, $\mathrm{DOR_{sim}}$ is the simulated DOR, and $\mathrm{inf_m}$ is the instantaneous monthly inflow into the reservoir in month, m. Also note that the ratio $\mathrm{DOR_{sim}}/\mathrm{DORT}$ is constrained to the interval [0, 1].

TSL and DORT were co-varied across the ranges TSL∈ {0.65, 0.75, 0.85, 0.95} and DORT∈ {0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00}. The default values are 0.85 and 0.5 for TSL and DORT respectively.

To understand the influence of accuracy of dam inflow simulations on *generic reservoir operation schemes* (GROS), sensitivity tests were performed on three specifically selected dams: John H. Kerr, Glen Canyon and Libby. John H. Kerr dam represents cases where H08 simulates inflow really well. This enables us to test whether the GROS correctly reproduces observed reservoir behaviour (*e.g*., tracking instantaneous inflow (run-of-the-river) vs. heavy regulation) without the confounding effect of inflow bias, and to determine if both sate (storage) and flux (release) variables can be optimised simultaneously. Conversely, Glen Canyon and Libby dams represent cases where H08 substantially overestimates or underestimates inflow respectively. This allows us to evaluate how the GROS dynamically responds to inflow errors and expose the potential trade-offs between optimising storage and release when simulated inflow is biased. By systematically evaluating model performance under each parameter pair for these contrasting scenarios, we can better understand the limits and physical implications of parameter tuning.

Meanwhile, for H08-LIS, we examined the *normal storage limit* (LN), a key LIS parameter that defines the storage fraction at which the reservoir switches between two primary operating modes:

- Storage conservation mode (storage fraction (F) < LN), where releases are reduced to ensure that storage is recovered towards LN.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

- Release-dominant mode (F > LN), where releases increase to draw the reservoir back towards LN, reflecting post-flood storage management behaviour. The basic release algorithm for LIS is shown in Eq. (S3).

To assess the sensitivity of LIS to this threshold (LN), we varied LN across the range {0.25, 0.50, 0.75}. Note that the default parameter value of LN for CWT runs in ISIMIP3a is 0.50.

3. Results

3.1. Validation of satellite-derived storage products

This section describes the intercomparison of satellite-derived storage datasets against *in situ* storage. The goal is to evaluate their appropriateness as surrogates for *in situ* storage and to identify the best performing satellite product to serve as a benchmark for evaluating simulated storage from GHMs.

To evaluate the performance of satellite-derived storage products, we compare them with the *in situ* observations (Fig. 2 and Table 3). Across the 638-reservoir validation sub-set (Sect. 2.3.1–2.3.2 and Fig. S1 in Sect. S2 of the Supplementary), the Global Reservoir Storage (GRS) demonstrated the strongest agreement with *in situ* storage (Fig. 2 and S2). It achieved a median Kling-Gupta Efficiency (KGE) and Symmetric Mean Absolute Percentage Error (SMAPE) of 0.44 and 25% respectively (Fig. 2a and b). Furthermore, GRS exhibited the narrowest *interquartile range* (IQR) and the highest fraction of reservoirs meeting performance thresholds, with 93% of reservoirs having a skilful KGE (KGE > -0.41) and 79% having a SMAPE below our self-imposed threshold (50%). In contrast, Glolakes and Global Dam Storage (GDS) exhibited lower median KGE and higher median SMAPE, alongside broader distributions in these skill metrics. They also yielded smaller fractions of reservoirs meeting the skill thresholds (GloLakes: 81% and 55% for KGE and SMAPE respectively; GDS: 88% and 63% for KGE and SMAPE respectively, with GDS outperforming GloLakes). These aggregate statistical differences are visually evident in the time series of individual reservoirs (Fig. 2c–f). Specifically, while the overall temporal dynamics are often well captured by the three products, their absolute volume estimates can deviate severely. For example, GloLakes exhibits massive storage overestimations at Libby and Rana Pratap Sagar dams (Fig. 2c and e), and similarly, GDS shows a large positive bias at Ebro dam (Fig. 2d). In contrast, GRS consistently tracks the *in situ* storage volumes across these distinct locations without extreme bias. This robustness is further supported by an analysis of the individual KGE components (Fig. S2). Although Glolakes exhibited a median bias term (β) closest to the optimal value of 1, and the highest Pearson's correlation (R) (Fig. S2b and c), GRS demonstrated less variance for each skill score, as evidenced by its narrower IQRs for these KGE terms. The superior performance of GRS over GDS, despite both relying on Global Reservoir Surface Area Dataset (GRSAD) area time series, highlights the critical role of reservoir bathymetry in retrieving reliable storage estimates. Consequently, motivated by its strong statistical performance and comprehensive global coverage (GRS explicitly accounts for 7,245 dams, covering the vast majority of the 7,320 records in the GRanD v1.3 (Lehner *et al*. 2011) database), we adopted GRS as the observation benchmark for our GHM evaluation.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

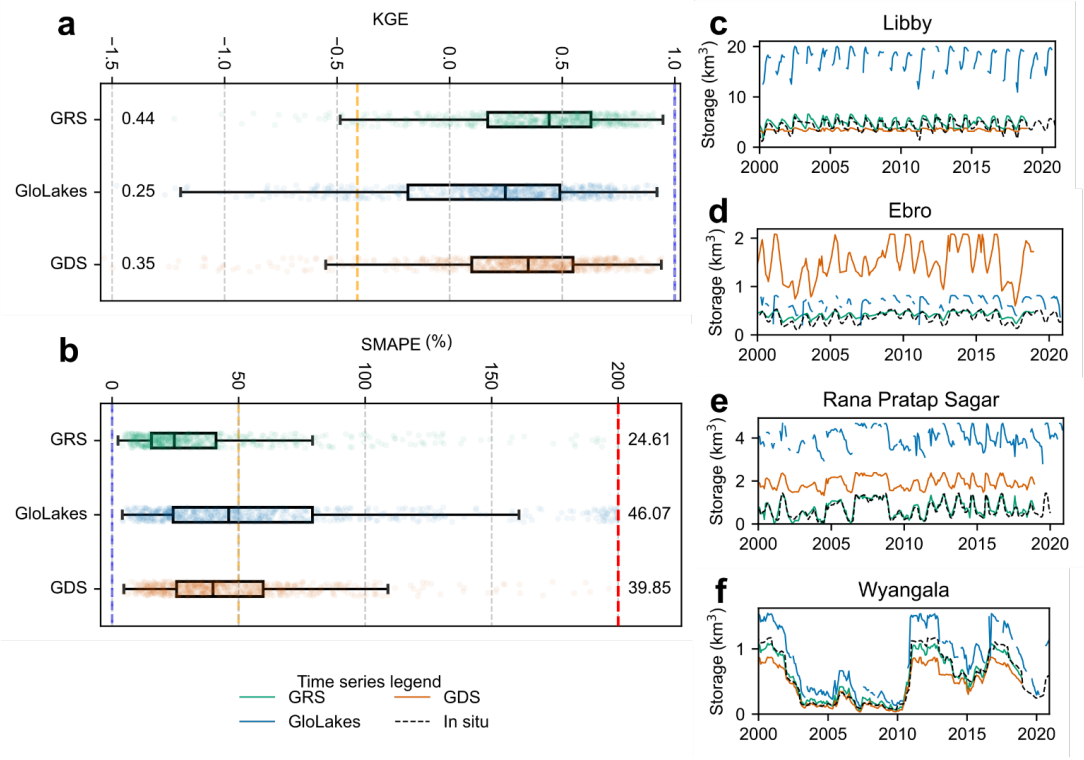


Figure 2: Performance evaluation of satellite-based reservoir storage products (N = 638 dams) against *in situ* storage. (a) Box plots of KGE scores for GRS, GloLakes and GDS, with overlaid strip plots coloured according to the time series legend. The dashed orange line marks the threshold for skilful KGE (-0.41) while the dashed blue line indicates the optimal KGE value (1). Median values are shown as vertical black lines within each box plot, with corresponding numerical labels. Whiskers represent 1.5× IQR, where IQR is the interquartile range. (b) Same as panel (a), but for SMAPE. The dashed red line indicates the worst possible SMAPE value (200%). (c–f) Representative long-term monthly series (2000–2020) for selected dams.

Table 3: Evaluation of GRS, GloLakes and GDS against *in situ* storage observations for selected dams using KGE and its components.

GRanD ID	Dam Name	Metric	Satellite Datasets		
			GRS	GloLakes	GDS
297	Libby	KGE	0.69	-2.08	0.11
		R	0.78	0.95	0.68
		α	0.83	1.82	0.18
		β	1.13	3.97	0.86
2662	Ebro	KGE	0.51	0.09	-2.78
		R	0.90	0.68	0.69
		α	0.55	1.31	3.36
		β	1.17	1.79	3.94
4836	Rana Pratap Sagar	KGE	0.90	-3.85	-0.85
		R	0.94	0.87	0.95
		α	0.99	1.16	0.73
		β	1.08	5.85	2.83

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

GRanD ID	Dam Name	Metric	Satellite Datasets		
6605	Wyangala	KGE	0.89	0.58	0.63
		R	0.99	0.96	0.98
		α	0.89	1.18	0.72
		β	0.98	1.37	0.76

3.2. Global evaluation of *global hydrological model* (GHM) storage simulations

To evaluate the performance of the storage estimations of ISIMIP3a GHMs, we calculated the KGE (Fig. 3 and Table 4). Across the 424 dams available in all datasets (GHM simulations and GRS), the median KGE for three out of five GHMs—H08, CWatM (CWT) and LPJml5-7-10-fire (LPJ)—fell below the established skill threshold of -0.41, while the remaining two models, MIROC-INTEG-LAND (MIL) and WaterGAP2-2e (WGP), performed marginally above the threshold with KGEs of -0.36 and -0.37 respectively (Fig. 3a). The proportion of skilful reservoirs (KGE > -0.41) was: MIL 52.36%, WGP 52.12%, H08 40.80%, CWT 40.33%, and LPJ 30.42%. The consistently low median values and low proportion of skilful dams indicate that generally, all five GHMs exhibit limited skill when benchmarked against GRS. Time series diagnostics at four selected dams—Fig. 3b–e and Fig. S8—revealed a consistent pattern whereby H08, WGP, LPJ and CWT simulated exaggerated seasonal drawdown (over-amplified intra-annual variability), while MIL produced damped storage cycles with reduced intra-annual variability relative to observations. Supporting performance metrics are summarised in Table 4. Crucially, these findings remain robust across the different climate forcings (Fig. S13–S15).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

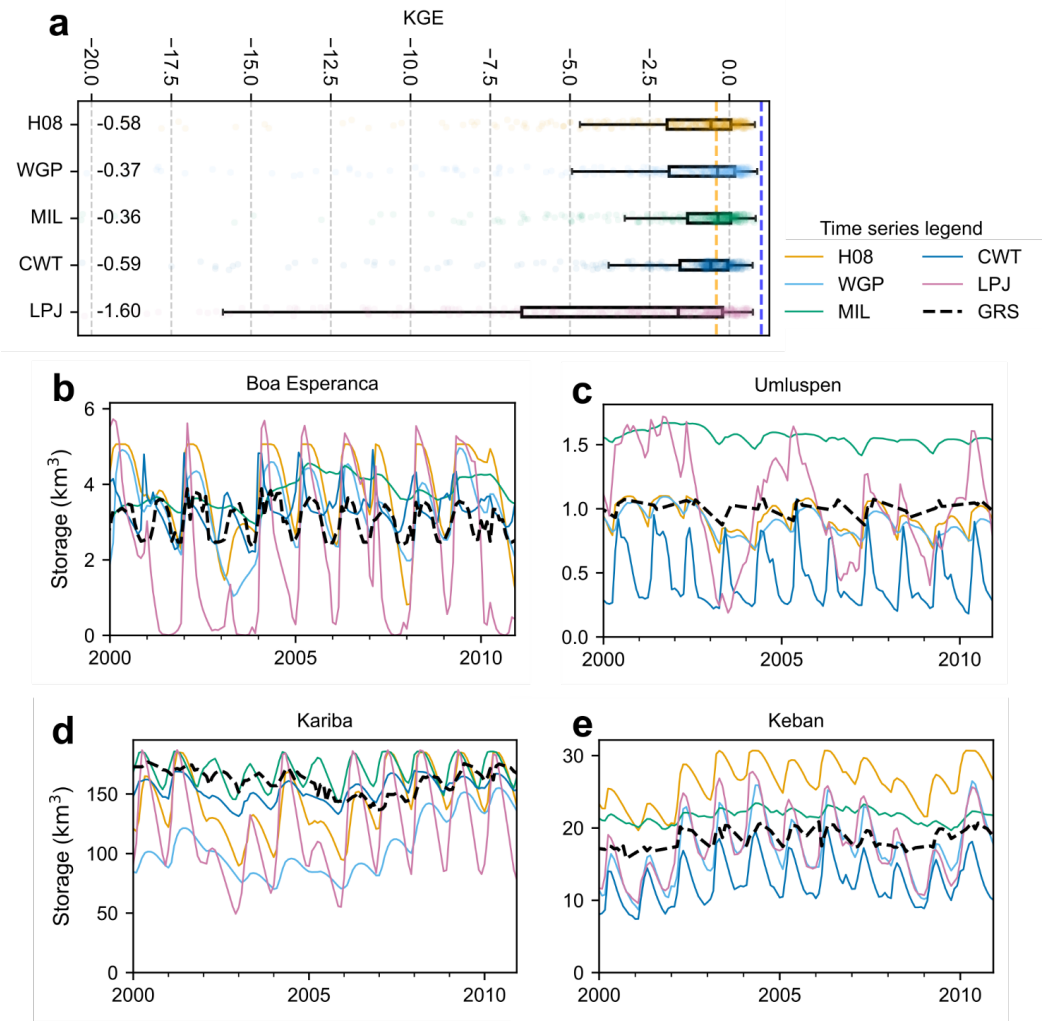


Figure 3: Performance evaluation of reservoir storage simulations by ISIMIP3a GHMs against GRS (N = 424 dams). (a) Box plots of KGE scores for five ISIMIP3a GHMs (H08, WGP, MIL, CWT and LPJ) with overlaid strip plots coloured according to the time series legend. The vertical, dashed orange line marks the threshold for skilful simulations (KGE = -0.41), while the vertical, dashed blue line indicates the best possible value (KGE = 1). Median values are denoted by vertical, black lines within each box plot, with corresponding numerical annotations. Whiskers represent 1.5 × IQR. (b–e) Representative long-term monthly series (2000–2010) for select dams.

Table 4: Performance metrics—KGE scores and their corresponding correlation components (α, β and R terms)—for the five ISIMIP3a GHMs across selected dams. Note that these metrics are derived from long-term monthly time series without bias correction.

GRanD ID	Dam Name	Metric	ISIMIP3a GHMs				
			H08	WGP	MIL	CWT	LPJ
2490	Boa Esperança	KGE	-0.71	-0.34	0.07	-0.06	-2.40
		R	0.58	0.46	0.10	0.09	0.34
		α	2.65	2.23	1.12	1.53	4.32
		β	1.23	1.04	1.22	1.10	0.68
3720	Umluspen	KGE	-0.52	-0.30	0.09	-2.69	-6.44

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

GRanD ID	Dam Name	Metric	ISIMIP3a GHMs				
		R	0.32	0.34	0.26	0.24	0.19
		α	2.36	2.11	1.00	4.57	8.39
		β	0.92	0.88	1.54	0.44	1.06
4056	Kariba	KGE	-0.39	-0.22	0.22	0.43	-1.68
		R	0.53	0.52	0.22	0.47	0.36
		α	2.31	2.07	1.04	0.80	3.59
		β	0.89	0.68	1.06	0.95	0.73
4422	Keban	KGE	-0.57	-1.19	0.25	-0.42	-1.52
		R	0.64	0.71	0.30	0.55	0.58
		α	2.47	3.16	0.77	2.30	3.48
		β	1.39	0.90	1.15	0.65	0.97

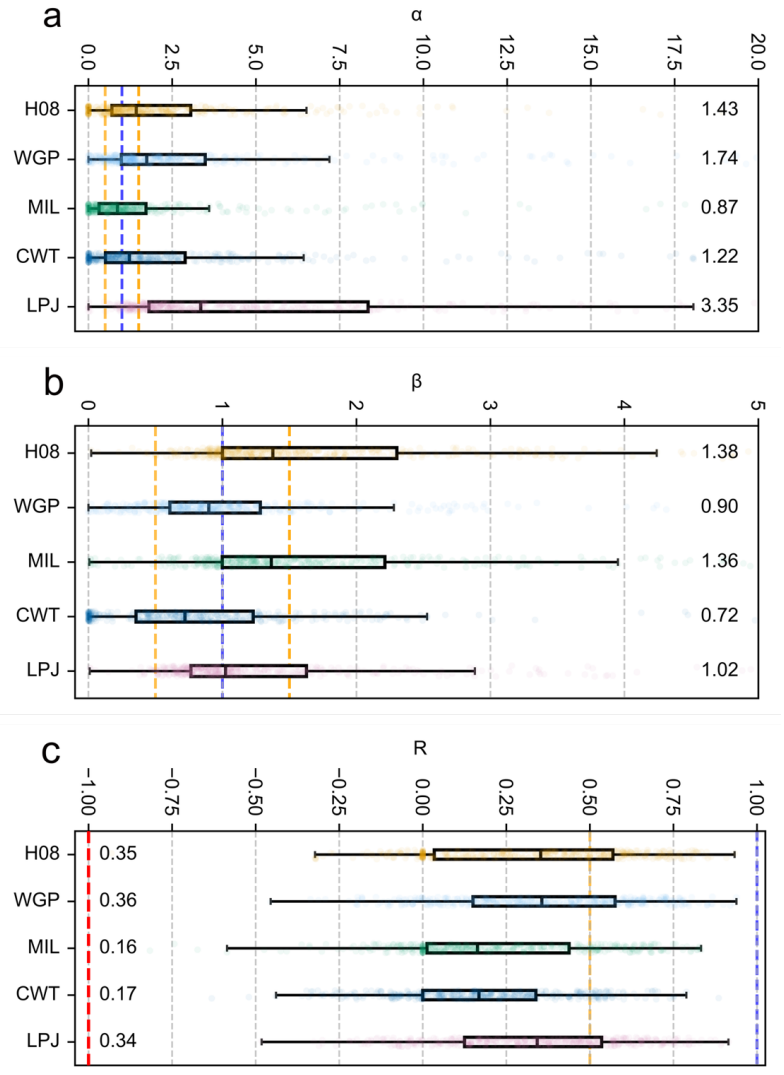
To test the sensitivity of our results to the satellite-derived observational benchmark, we repeated the GHM evaluation using the GDS dataset. Although the absolute storage volumes differ between GRS and GDS (Sect. 3.1), the fundamental evaluation of the GHMs remains robust (Fig. S17 and S18). While the ordinal ranking of the top two performing models fluctuates depending on the dataset (MIL and WGP swap first place (Fig. 3 and S17)), the overall performance patterns remain highly consistent. This consistency not only underscores the utility of the GDS product but also confirms that our broader diagnostic findings regarding the GHMs are robust to the choice of either GRS or GDS as an observation benchmark.

Furthermore, to ensure our evaluation was not confounded by the structural diversity in how GHMs classify reservoirs (*i.e.*, applying complex reservoir operating schemes to *global* reservoirs vs. simpler modelling of *local* ones), we conducted an additional analysis. We re-evaluated model performance using a strictly controlled subset of 321 dams that are universally classified as *global* across H08, WGP and MIL. As shown in Fig. S19, the overall KGE scores and relative model rankings for this universally *global* dams subset are highly consistent with the results from the full 424-dam ensemble (Fig. 3a). This confirms that the observed performance deficits are driven fundamentally by the active reservoir operating schemes (H06 or LIS) themselves, rather than being a penalty for models which simulate a larger fraction of these dams as simple natural lakes.

3.3. Decomposition of simulation errors (the components of KGE)

Before attempting any corrections, we first needed to isolate the specific causes of model error in the raw simulations. To identify the dominant driver of simulated storage error, we performed KGE decomposition (Fig. 4 and Table 4). KGE decomposition showed that the variability term (α) was the dominant contributor to error for most GHMs (Fig. 4). H08, WGP, LPJ and CWT overestimated seasonal amplitude: median $\alpha \in [1.22, 3.35]$. On the other hand, MIL exhibited a median α of 0.87, closer to observed variability. However, MIL also showed comparatively low temporal correlation (median R = 0.16) and a larger mean bias among the five models (Fig. 4). Within the group with exaggerated amplitude, H06-based models—H08, WGP, LPJ—attained higher temporal correlations (median $R \in [0.34, 0.36]$) than CWT (median R = 0.17), despite similar or better mean bias in CWT (Fig. 4 and Table 4). Importantly, just as with the overall KGE scores, this dominance of amplitude and mean bias errors persists even when strictly evaluating the subset of 321 dams that are universally classified as *global* dams by all GHMs (Fig. S20).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026



425 **Figure 4: Same as Fig. 3a, but for the KGE components: the alpha (α), beta (β) and the correlation (R) terms.**

The controlled comparison between H08-H06 and H08-LIS (Fig. S7) reveals that the choice of reservoir operation scheme induces distinct performance characteristics. H08-H06 achieved a slightly higher median temporal correlation (R = 0.35) compared to H08-LIS (R = 0.31) (Fig.S7c), suggesting that the process-based logic of H06 may offer a modest advantage in capturing release timing over the piecewise-function logic of LIS. However, H08-LIS achieved a better overall median KGE
430 (-0.38 vs. -0.58: Fig. S7a), largely because H08-H06 suffers from a substantially larger mean-storage bias (median β=1.37 vs. 0.89 for H08-H06 and H08-LIS respectively) with its default global parameter TSL = 0.85 (Fig. S7d), which causes reservoirs to be operated with too high a storage.

3.4. Impact of targeted statistical bias correction

Having identified variability and mean bias as the primary culprits in Sect. 3.3, we next quantified exactly how much these
435 specific flaws were degrading overall model skill. As a diagnostic baseline, we applied the targeted statistical transformations described in Sect. 2.4.3 to see if artificially curing these errors would recover model skill. To isolate the influence of amplitude

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

errors, we applied a variance-matching bias correction that forces $\alpha \approx 1$ while intentionally preserving the simulated mean, to conserve as much of the inherent model characteristics as possible. This adjustment increased the proportion of skilful reservoirs to 87.04% (WGP), 82.15% (LPJ), 80.68% (CWT), 69.19% (MIL) and 68.70% (H08), and altered the median KGE ranking to: WGP (0.19), LPJ (0.18), H08 (0.01), CWT (-0.04), and MIL (-0.04) (Fig. 5 and Table 5). The reversal of comparative performance confirms that MIL's apparent skill in the original simulations (simulations without bias correction) is primarily a statistical artefact of dampened variability, whereas the other GHMs are heavily penalised for exaggerated amplitude rather than deficiencies in temporal correlation. Furthermore, skill recovery after linear bias correction—forcing $\beta = 1$—was particularly pronounced for MIL (improving the skilful rate from 52.36 to 73.08% of dams) and H08 (40.80 to 60.09 %), while comparatively minimal for WGP, CWT and LPJ (Fig. S11 and S12 and Table S1). It is then apparent that H08 and MIL simulate global reservoirs as operating too full, overestimating global water stored behind dams. It is worth noting that liner bias correction carries an inherent constraint, arising from the fact that linear scaling rigidly couples the adjustment of the mean and that of the variance. Consequently, linear bias correction is highly effective when a model's mean bias (β) and variability error (α) are directionally aligned: *i.e.*, when a model overestimates both mean storage volume and seasonal amplitude or vice versa. When errors are mixed, linear scaling forces a mathematical trade-off. For models like MIL, eliminating the large mean biases often improved the overall KGE score, even when it further reduced their already accurate or underestimated variability. However, for models that underestimate the mean but exaggerate the amplitude—this tendency is evident in the median scores for WGP and showcased by WGP at Keban dam or LPJ at Boa Esperança dam (Table 4)—scaling the mean up to match observations (forcing $\beta \approx 1$) requires multiplying the time series by a factor > 1. This inadvertently inflates the already exaggerated variability further (Table S1). In these cases, the heavy KGE penalties for massive amplitude errors easily outweigh the benefits of correcting the mean (Table S1).

Nonetheless, collectively, these results indicate that correcting amplitude and generally mean bias yields substantial performance gains across GHMs and alters their relative performance rankings. The widespread recovery of model skill achieved through the more targeted variance-matching adjustment confirms that exaggerated intra-annual variability is the dominant error degrading model performance.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

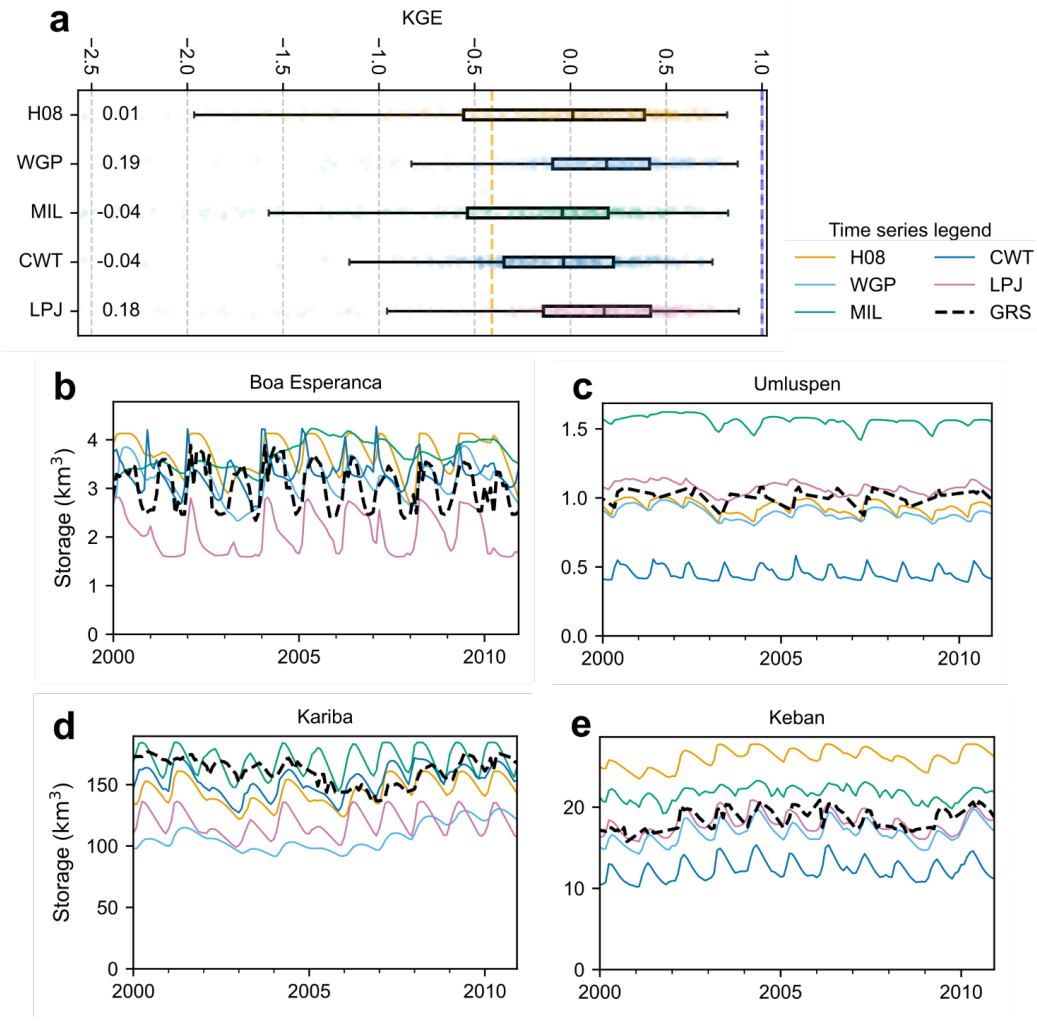


Figure 5: Same as Fig. 3, but evaluating the storage simulations after applying variance-matching bias correction.

Table 5: Same as Table 4, but evaluating the storage simulations after variance-matching bias correction.

GRanD ID	Dam Name	Metric	ISIMIP3a GHMs				
			H08	WGP	MIL	CWT	LPJ
2490	Boa Esperança	KGE	0.52	0.45	0.02	0.08	0.27
		R	0.58	0.47	0.07	0.08	0.35
		α	0.92	0.87	0.75	0.94	0.92
		β	1.23	1.04	1.22	1.10	0.68
3720	Umluspen	KGE	0.32	0.33	0.11	0.06	0.19
		R	0.32	0.34	0.32	0.24	0.19
		α	0.97	0.96	0.79	0.98	0.92
		β	0.92	0.88	1.54	0.44	1.06
4056	Kariba	KGE	0.54	0.41	0.21	0.46	0.31

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

GRanD ID	Dam Name	Metric	ISIMIP3a GHMs				
		R	0.56	0.51	0.21	0.47	0.36
		α	0.95	0.94	0.96	0.93	0.97
		β	0.89	0.68	1.06	0.95	0.73
4422	Keban	KGE	0.45	0.69	0.27	0.42	0.56
		R	0.61	0.71	0.32	0.53	0.56
		α	0.97	0.99	0.80	0.97	0.97
		β	1.39	0.90	1.15	0.65	0.97

3.5. Parameter sensitivity test of reservoir operating schemes

While the statistical bias corrections in Sect. 3.4 successfully demonstrated that resolving amplitude and mean bias significantly improves overall model skill, these adjustments remain artificial. To determine if these improvements could be achieved mechanistically, we investigated whether targeted parameter tuning could naturally resolve the observed errors. The five GHMs employ either the H06 (H08, WGP, MIL and LPJ) or LIS (CWT) reservoir operating schemes, which traditionally rely on uniform, fixed parameters for global simulations. By conducting sensitivity tests on these models' parameters for three selected dams (John H. Kerr, Glen Canyon and Libby), we aimed to determine if adjusting the modelled reservoir operating rules could replicate the success of our diagnostic statistical corrections.

3.5.1 H06 Parameter sensitivity test

As has been established, John H. Kerr dam presents a unique case where H08 simulated inflow into the dam quite well: Fig. 6a shows simulated inflow tracking *in situ* observed inflow well, with a strong performance in KGE and all its components. Increasing DORT from the default 0.50 to 2.00, while keeping TSL at 0.85, improved storage simulations markedly, raising the KGE from -1.80 to -0.31 (Fig. 6b and d), achieving a skilful score (KGE > -0.41). This indicates that the default, DORT = 0.5, wrongly classified the dam as heavily regulated (both in simulation and real-world scenarios: see Table 6 for observed and simulated DOR), whereas a higher DORT correctly reflected its release operation as heavily dominated by instantaneous monthly inflow (~ 83% as shown in Table 6) and to a smaller degree by the long-term annual discharge ($0.17 \times$ Krls). Similarly, for dam release, the best performance was achieved with DORT = 2.00 and TSL = 0.85, increasing KGE from 0.49 to 0.66 (Fig. 6c and e). Notably, a single parameter set optimises both storage and release simultaneously: both variables (storage and release) achieve peak skill at TSL = 0.85 and DORT = 2.00. This consistency is expected, first, because H08 accurately simulates inflow for the John H. Kerr dam and, second, because the H06 release for this particular dam is dominated by this instantaneous monthly inflow (Fig. 6d–e and Fig. 7).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

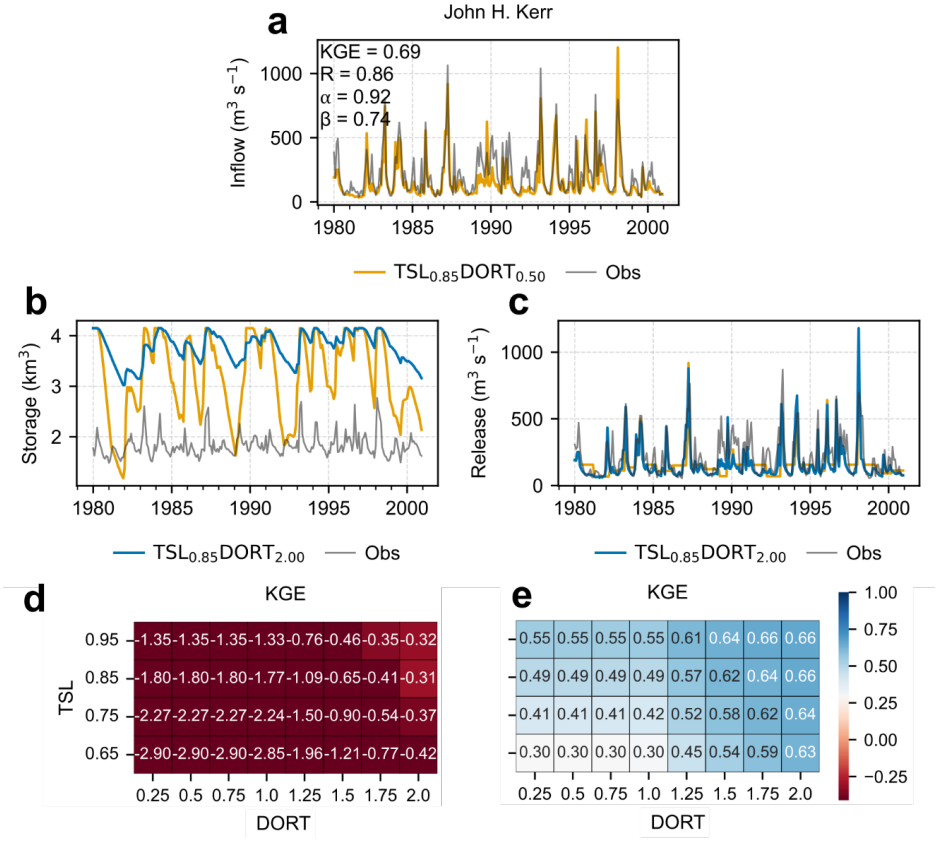


Figure 6: Sensitivity analysis of H06 parameters at John H. Kerr dam. (a) Long-term monthly observed (*in situ*) inflow compared to H08 simulations using default global parameters (TSL = 0.85 and DORT = 0.5). (b) Comparison of observed (*in situ*) storage against H08 storage simulated with default global parameters (vermilion curve) and the best-performing tuned parameter set (blue curve). (c) Same as panel (b) but for release time series. (d) Storage KGE heat map for the parameter sensitivity test annotated with KGE scores. (e) Same as panel (d) but for release (outflow) time series.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

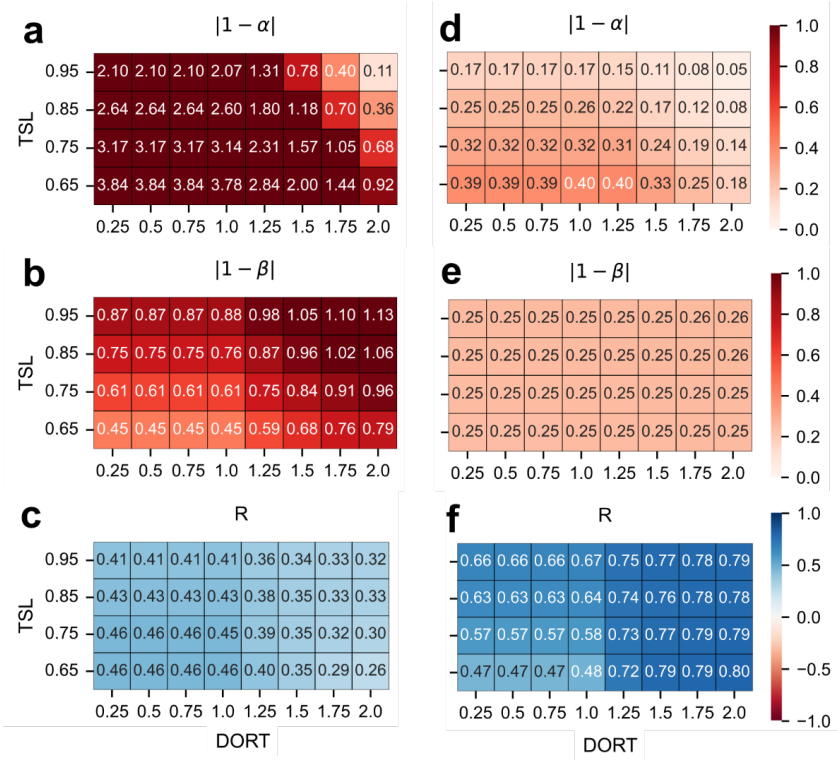


Figure 7: Same description as Fig. 6d–e, but the components of KGE. Note that for the alpha (α) and beta (β) terms, the heat map annotations shown are the absolute values of the deviation from one.

Table 6: Impact of parameter DORT and the *degree of regulation* (DOR) on the H06 release formulation for selected dams. The release is calculated as $R_{ls} = wR_{target} + (1 - w)I_m$, where the weight, $w = (DOR_{sim}/DORT)^2$. This is simply a reformulation of Eq. (11). Note that DOR_{obs} and DOR_{sim} are the corresponding DORs for observed and simulated long-term annual dam inflows. Also note that $w \in [0, 1]$.

Dam Name	DOR_{obs}	DOR_{sim}	DORT	Weight (w)	Dominant Factor	Simulated Operation Mode
John H. Kerr	0.61	0.82	0.50	1.00	Target Release	Heavily regulated (no influence of monthly inflow at all)
			2.00	0.17	Monthly Inflow	Inflow dominated (minimal influence of target release)
Glen Canyon	1.70	0.89	0.50	1.00	Target Release	Heavily regulated (no influence of monthly inflow at all)
			1.00	0.79	Target Release	Moderately heavily regulated (minimal influence of monthly inflow)
			1.25	0.50	Balanced	Balanced (monthly inflow and target release have almost equal influence)
Libby	0.75	0.89	0.50	1.00	Target release	Heavily regulated (no influence of monthly inflow at all)
			0.75	1.00	Target Release	Heavily regulated (no influence of monthly inflow at all)

Conversely, Glen Canyon dam provides a typical case of dam inflow simulation with a significant positive bias (over estimation). This can be visualised in Fig. 8a (β=1.91) and from Table 6, where $DOR_{sim} < DOR_{obs}$, meaning that in reality, the dam has a greater carry-over-capacity than simulation suggests. Increasing DORT from 0.50 to 1.25, while keeping TSL at 0.85, improved storage KGE from 0.59 to 0.66 (Fig. 8b and d). Despite H08 overestimating inflow, the H06 scheme effectively throttled releases, shifting the dam's operational mode away from a strictly heavily regulated system (as classified with default

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

global parameters) toward a balanced mode: a mid-point between a heavily regulated dam and one where release tracks instantaneous monthly inflow). Refer to Table 6 for the mathematical concept (weighting changes in the release equation). For release, DORT = 1.00 and TSL = 0.85 yielded the best performance, improving KGE from -0.06 to -0.04. Here, the release decision is dominated by fixed annual flow (0.79 × Krls), with monthly inflow contributing about 21%, *i.e.*, release simulation is best when the dam release logic approaches the heavy regulation mode. Critically, the optimal DORT values for storage and release differ significantly, underscoring the trade-off between optimising state (storage) and flux (release) variables when inflow is over-estimated.

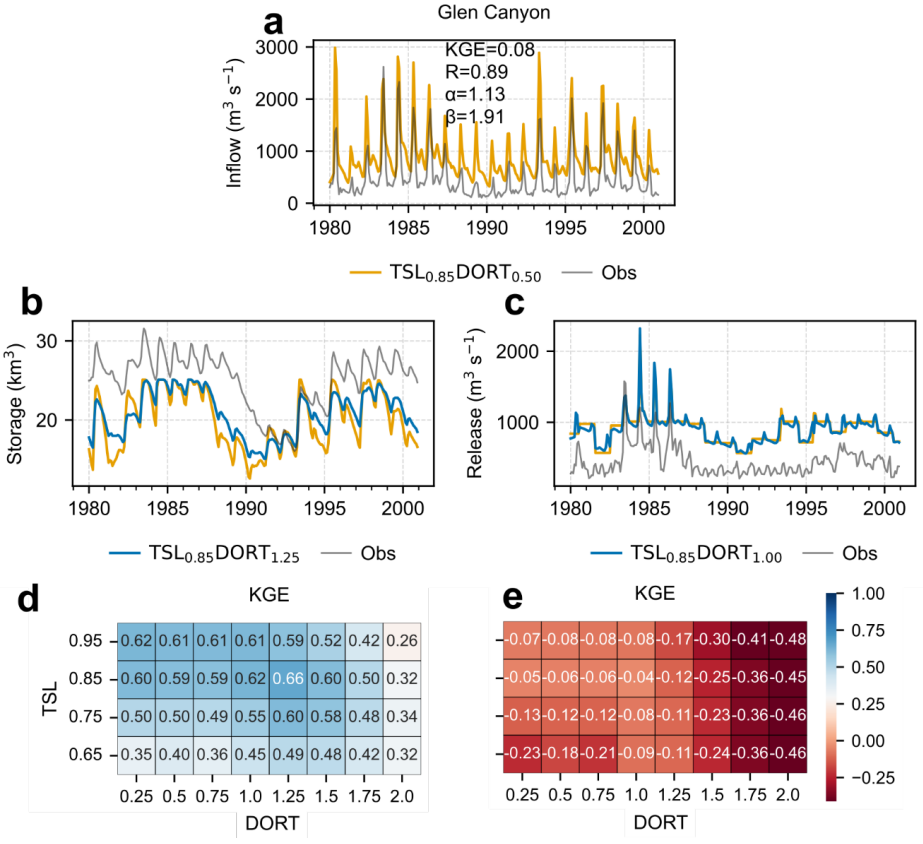
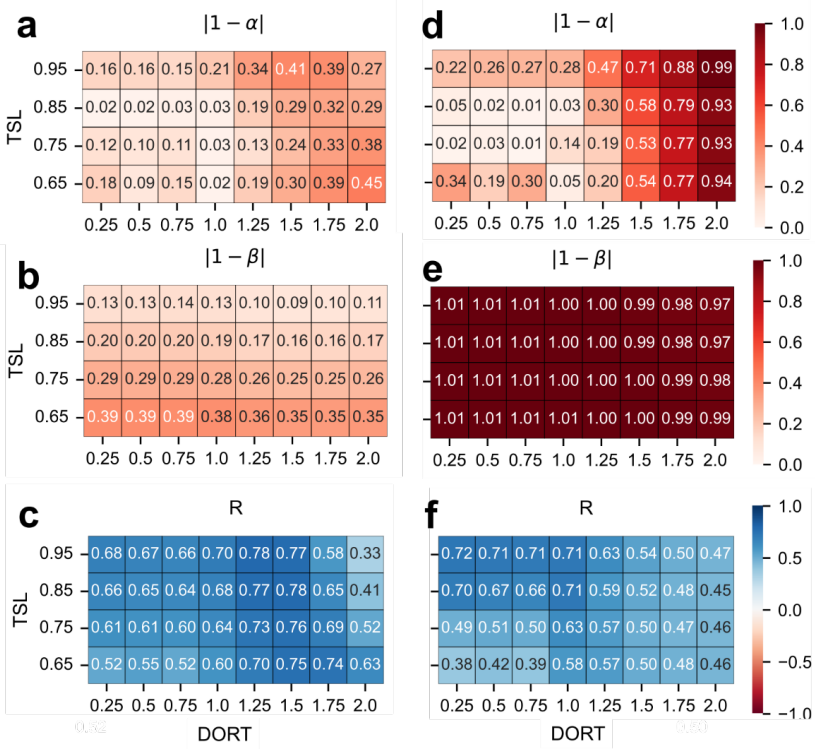


Figure 8: Same as Fig. 6, but for Glen Canyon dam.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026



515 **Figure 9: Same as Fig. 7, but for Glen Canyon dam.**

Libby dam provides a case of underestimated dam inflow. This can be visualised in Fig. 10a (β = 0.84) and from Table 6, where $DOR_{sim} > DOR_{obs}$, meaning that in reality, the dam has a smaller carry-over-capacity than simulation suggests. DORTs of 0.25, 0.50 or 0.75 and decreasing TSL to 0.65 improved the storage KGE from 0.11 to a peak of 0.34 (Fig. 10b and d). For release, DORTs of 0.25, 0.50 or 0.75 and TSL = 0.85 yielded the best performance, similar to the default parameter set. Here,
520 the ideal release decision is set to fixed annual flow (1.00 × Krls). Critically, while DORT values agree, the optimal TSL values for storage and release differ, again, underscoring the trade-off between optimising state (storage) and flux (release) variables. Together, these tests show that discrete adjustment of H06 parameters can improve storage and/or release skill, though optimal settings may differ by site and by variable (state vs. flux). Figure 6–Fig. 11 and Table 6 summarise these outcomes.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

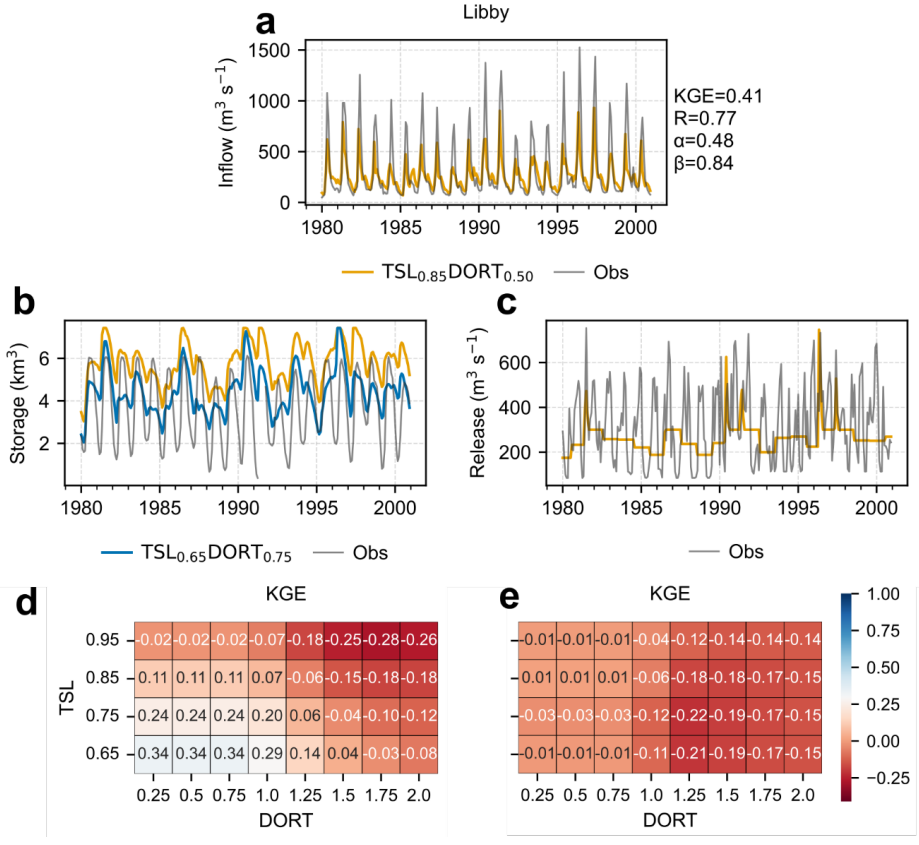


Figure 10: Same as Fig. 6, but for Libby dam.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

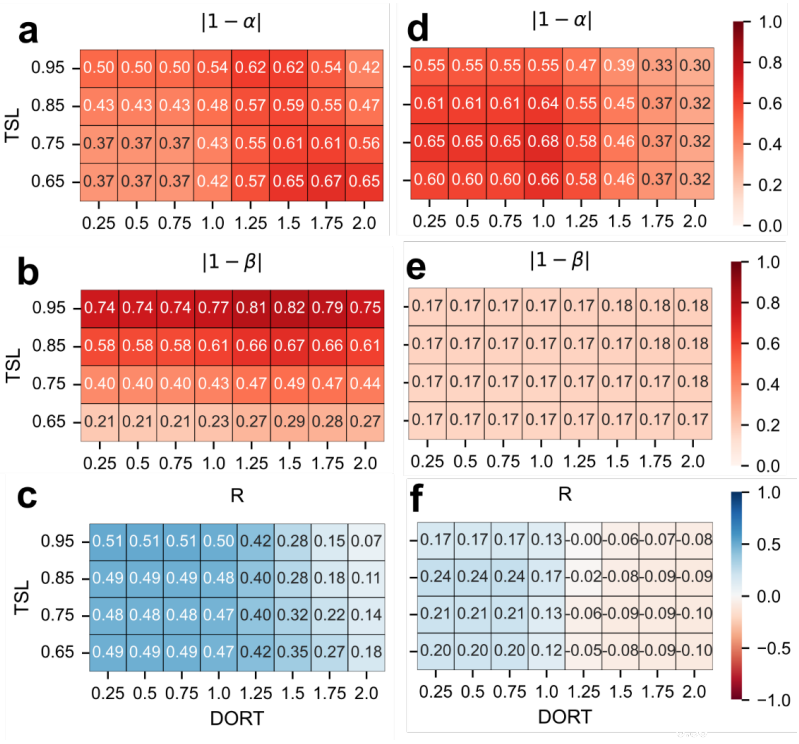


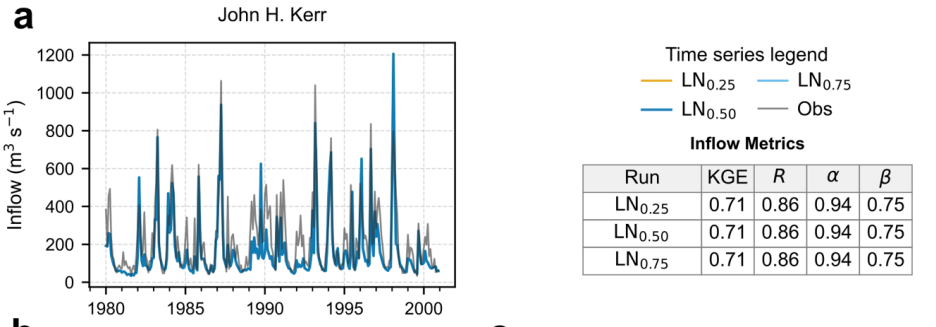
Figure 11: Same as Fig. 7, but for Libby dam.

To ensure our findings were not isolated to the three selected case study dams, identical parameter sensitivity tests were conducted on four additional dams representing diverse inflow and operational conditions: Bull Shoals, Copeton, Boysen, and Burrinjuck dams. The results, provided in the Sect. S10 of the Supplementary (Figs. S21 and S22), corroborate our primary findings: that discrete adjustments to DORT and TSL can enhance either storage or release performance, but the optimal parameter combinations remain dependent on the target variable being optimised for cases of overly biased inflow simulation.

3.5.2. LIS parameter sensitivity test

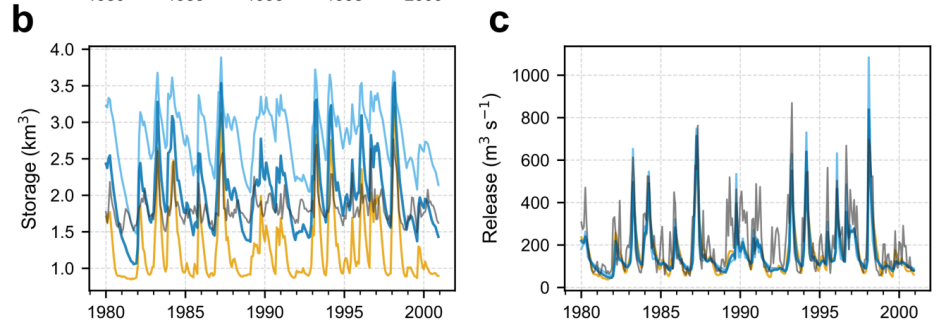
We further investigated whether the performance of the LIS scheme is constrained by its parameterisation, specifically the *normal storage limit* (LN). In LIS, zones of operation of a reservoir (excluding conservation and flood mode) are determined relative to LN (Eq. (S3)). LN effectively sets the dam operator's target storage level: lower LN values represent operating strategies where dams are kept near empty (typical of flood control dams), whereas higher LN values correspond to reservoirs operated near full (Fig. 12b, 13b and 14b). For John H. Kerr dam, increasing LN yielded minimal gains in overall storage skill (KGE remained ~ -0.27 to -0.31) (Fig. 12b). However, while LN values of 0.50 and 0.75 produced similar composite KGE scores, the relative contributions of the KGE components differed significantly. For example, LN = 0.50 achieved the highest correlation (R = 0.72) and a bias term closer to the optimal value of 1 (β =1.10) but exhibited higher variability error (α = 2.24). Conversely, LN = 0.75 improved variability (α = 2.07) at the expense of correlation (R = 0.59) and bias (β = 1.53). Rather than point to a single optimal parameter, this trade-off highlights how similar overall KGE scores can mask fundamentally different operational behaviours, underscoring the importance of evaluating individual KGE components during calibration. Release simulations also improved for this dam, albeit modestly, with KGE rising from 0.56 (LN = 0.25) to 0.64 (LN = 0.75) (Fig. 10c).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026



Inflow Metrics

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	0.71	0.86	0.94	0.75
$LN_{0.50}$	0.71	0.86	0.94	0.75
$LN_{0.75}$	0.71	0.86	0.94	0.75



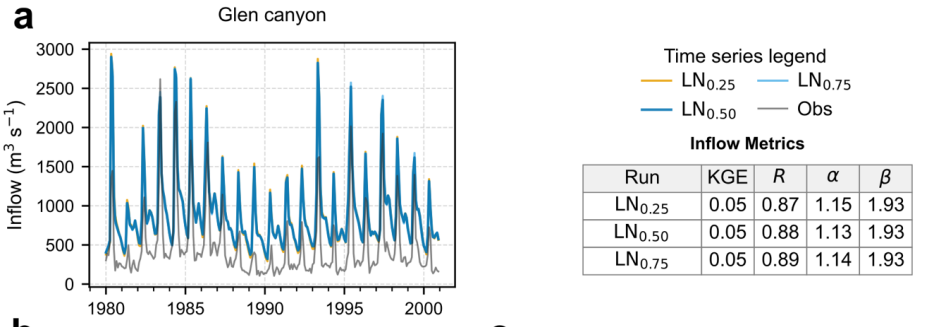
Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	-0.30	0.74	2.25	0.71
$LN_{0.50}$	-0.27	0.72	2.24	1.10
$LN_{0.75}$	-0.27	0.59	2.07	1.53

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	0.56	0.78	0.71	0.76
$LN_{0.50}$	0.59	0.78	0.75	0.76
$LN_{0.75}$	0.64	0.77	0.87	0.76

Figure 12: Sensitivity analysis of LIS' LN parameter at John H. Kerr dam. (a) Long-term monthly observed (*in situ*) inflow compared to simulations using LN = 0.50. (b) Comparison of observed (*in situ*) storage against storage simulated with different parameter sets. (c) Same as panel (b) but for release time series.

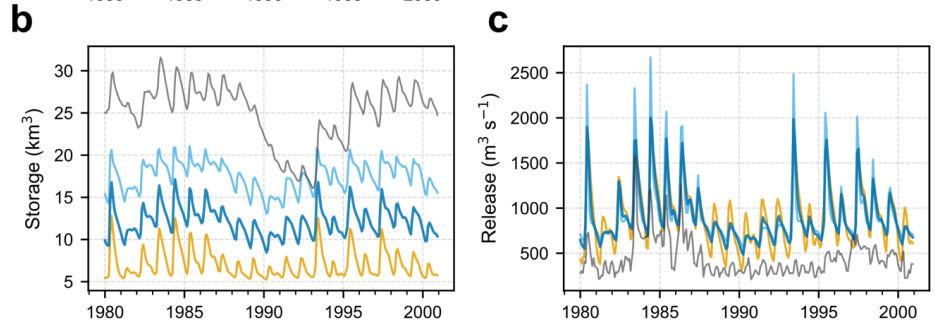
In contrast, Glen Canyon dam demonstrated high sensitivity to LN. Simulation with a low LN (0.25) resulted in poor storage performance (KGE = -0.06) (Fig. 13b). However, increasing the LN to 0.75, effectively maximising the amount of water impounded by the dam, significantly improved KGE to 0.28. For release, simulations with LN = 0.25 were poorest (KGE = -0.26) (Fig. 13c). The best performance was achieved with LN = 0.5 (KGE = -0.17).

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026



Inflow Metrics

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	0.05	0.87	1.15	1.93
$LN_{0.50}$	0.05	0.88	1.13	1.93
$LN_{0.75}$	0.05	0.89	1.14	1.93



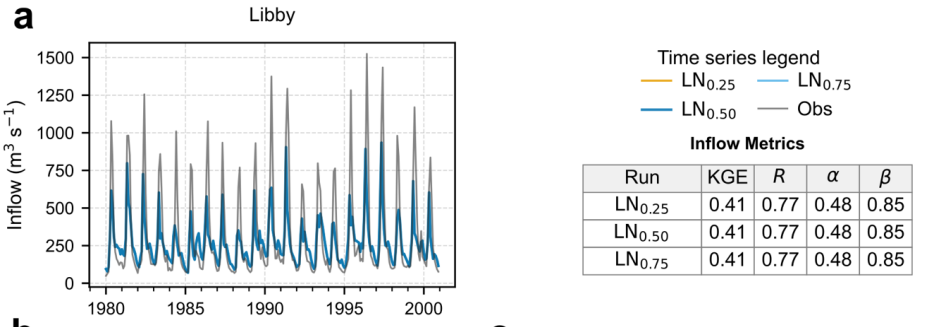
Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	-0.06	0.44	0.47	0.28
$LN_{0.50}$	0.13	0.52	0.49	0.48
$LN_{0.75}$	0.28	0.60	0.49	0.70

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	-0.26	0.58	1.60	2.03
$LN_{0.50}$	-0.17	0.66	1.44	2.03
$LN_{0.75}$	-0.23	0.67	1.59	2.03

Figure 13: Same as Fig. 12, but for Glen Canyon dam.

Libby dam also demonstrated high sensitivity to LN. Simulation with a low LN (0.25) resulted in poor storage performance (KGE = -0.05) (Fig. 14b). However, increasing the LN to the default 0.50 improved KGE to 0.16. Similarly, for release, simulations with LN = 0.25 were poorest (KGE = 0.06) (Fig. 14c). The best performance was achieved with LN = 0.5 (KGE = 0.10). Similar sensitivity to LN was observed across the four supplementary dams: Bull Shoals, Copeton, Boysen, and Burrinjuck (Fig. S23 and S24). Consistent with our primary case studies, these supplementary tests confirm that while tuning LN can yield substantial gains in either storage or release skill, a single uniform parameter set cannot universally optimise the storage and release in cases of biased inflow.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026



Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	0.41	0.77	0.48	0.85
$LN_{0.50}$	0.41	0.77	0.48	0.85
$LN_{0.75}$	0.41	0.77	0.48	0.85

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	-0.05	0.39	0.27	0.57
$LN_{0.50}$	0.16	0.53	0.31	0.98
$LN_{0.75}$	0.08	0.56	0.32	1.42

Run	KGE	*R*	*α*	*β*
$LN_{0.25}$	0.06	0.17	0.59	0.85
$LN_{0.50}$	0.10	0.26	0.51	0.85
$LN_{0.75}$	0.08	0.22	0.54	0.84

Figure 14: Same as Fig. 12, but for Libby dam.

4. Discussion

4.1. Mechanism underlying the variability-driven divergence among H06-based GHMs

Despite common use of the Hanasaki *et al*. (2006) scheme (H06) across H08, WaterGAP2-2e (WGP), LPJml5-7-10-fire (LPJ) and MIROC-INTEG-LAND (MIL), the models diverge markedly in their simulated variability. In the strictly controlled subset of 321 *global* dams, where all models apply the H06 scheme, H08, WGP and LPJ severely overestimate seasonal amplitude while MIL achieves a near-perfect median (α=1.04) (Fig. S20a). This divergence appears not to originate from the reservoir scheme itself, but from the quality of dam-inflows generated by the GHMs. From a global intercomparison of discharge simulations by ISIMIP3a GHMs, Heinicke *et al.* (2024) found that while H08 and WGP tended to better reproduce river discharge variability (median $\alpha \approx 1$), MIL substantially underestimated it (median $\alpha \approx 0.5$), capturing only about half of the observed variability. This indicates a distinct mechanism of compensating errors. When using the default parameters, the H06 scheme seems to be prone to exaggerating intra-annual storage variability. When models with more accurate runoff variability (*e.g.*, H08 and WGP) feed inflows into this default-parameter H06, the resulting storage amplitude becomes vastly exaggerated. For MIL, however, feeding inflows with a damped variability into the same scheme seems to act as a counterweight: the comparatively *flat* runoff effectively neutralises the default H06 algorithm's tendency to exaggerate corresponding storage amplitude, yielding an artificially accurate storage variability. This neutralisation effect is confirmed when evaluating the full 424-dam ensemble. MIL models about 17% of these dams as simple natural lakes, ignoring the H06 scheme entirely. With the H06 algorithm implemented in about 83% of the reservoirs, MIL's overall median KGE variability term then drops to 0.87 (Fig. 4a), and with H06 implemented across all reservoirs (universally *global* reservoirs), median variability rises to 1.04 (Fig. S20a). In other words, compared to other H06-based GHMs—specifically H08, which uses an identical runoff routing—MIL's

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

storage behaviour is not due to a structurally or parametrically different reservoir operation scheme (H06), but rather reflects potential limitations in its land surface runoff generation process. While this interpretation is preliminary and indicative, it offers a plausible explanation for MIL's consistently muted storage variability. A straightforward initial step to improve MIL's global storage simulations is to revise its capacity threshold for global reservoirs. By lowering its current 1.0 km³ threshold to 0.5 km³ (aligning with WGP), a greater proportion of reservoirs would be actively managed by the H06 algorithm. Given that the interaction between MIL's runoff and the H06 scheme currently yields highly accurate storage variability (α = 1.04), extending the scheme to these smaller reservoirs should effectively further correct the dampened variability (α = 0.87) observed in the broader ensemble, providing an immediate pathway to recover overall KGE skill.

4.2. Structural considerations of reservoir operation schemes

The comparison between H06-based GHMs—except MIL—and CWatM (CWT) shows that the former achieves consistently higher temporal correlation in storage simulations (Sect. 3.3). At first glance, the comparatively low temporal correlation of CWT might suggest a structural limitation in the reservoir operation scheme of Burek *et al*. (2013, 2020), LIS. However, our controlled experiment—implementing LIS within H08 (H08-LIS)—demonstrates that LIS can achieve temporal correlations comparable to H06 when embedded within the same host GHM (Fig. S7). Furthermore, the sensitivity analysis shows that LIS performance is highly dependent on the *normal storage limit* (LN) parameter, with its value influencing whether the reservoir operation scheme behaves in a storage-conserving or release-dominant mode. The underperformance of CWT in storage temporal correlation still remains puzzling in light of its good discharge performance reported by Heinicke *et al*. (2024). While speculative, a plausible explanation is parametric rigidity in the global simulation set-up—*i.e.*, the use of uniform, non-calibrated reservoir parameters across all reservoirs—which could suppress the LIS' ability to represent reservoir operations accurately.

4.3. Parametric rigidity

The sensitivity experiments indicate that errors attributed to the reservoir operation schemes—H06 or LIS—are largely parametric: altering the *degree of regulation threshold* (DORT), and the *target storage level* (TSL) (for H06) and LN (for LIS) produced material gains in storage and/or release skill at both test dams (Sect. 3.5). At the same time, the optimal parameter sets for storage (state) and release (flux) are sometimes dissimilar, indicating an inherent multi-objective trade-off. Fundamentally, this trade-off is heavily constrained by the mass balance of the reservoir. Because the change in storage is dictated by the difference between inflow and release, bias in the simulated upstream inflow mathematically forces the model to propagate the error into either the storage state or the downstream release flux. Previous studies have reported similar findings. For example, Yassin *et al*. (2019) developed the Dynamically Zoned Target Release (DZTR) scheme and demonstrated, through multi-objective optimisation, that distinct trade-offs exist between reproducing reservoir storage and downstream release, implying that a single parameter set rarely maximises performance for both state and flux variables simultaneously. Döll *et al*. (2024) demonstrated that calibrating WGP against river discharge alone degraded the simulation of Total Water Storage Anomalies (TWSA) in the Mississippi River basin, whereas simultaneous assimilation of both discharge and NASA's Gravity Recovery and Climate Experiment (GRACE) satellite observations was required to balance the trade-offs between reproducing state (storage) and flux (discharge) variables. In the Ganges and Brahmaputra basins, Hasan *et al*. (2025) utilised a multi-objective evolutionary algorithm to demonstrate that calibrating WGP solely against river discharge degraded the simulation of TWSA and evapotranspiration. They concluded that simultaneous assimilation of multiple variables is necessary to mitigate these structural trade-offs, even if it prevents achieving the global optimum for any single variable. Finally, Hosseini-Moghari and Döll (2025) demonstrated that while assimilating satellite-derived storage anomalies generally improved the release performance of WGP, there remained a subset of reservoirs where forcing the model to match observed

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

storage dynamics led to a deterioration in downstream release simulations, highlighting the structural inconsistencies and mass balance limitations that might arise when optimising for a single variable.

4.4. Limitations of the study

Our evaluation relies on satellite-derived storage, specifically the GRS dataset, as a surrogate for *in situ* storage. As with any satellite-derived product, uncertainties arise, specifically from the detection of open surface water extent and from area-to-volume conversion where bathymetry is approximated: for example, Li *et al*. (2023) adopted standard geometric approximations for 6,898/7,245 reservoirs in the GRanD v1.3 inventory to create GRS. These uncertainties introduce noise and potential bias relative to ground (*in situ*) observations, evidenced by a mean Normalised Root Mean Square Error (NRMSE) of 24.30% for the simulated geometric method compared to 13.28% for observed bathymetry-based estimates (Li *et al*. 2023), which inevitably propagates into our model evaluation. At the same time, rapid advances in satellite-based reservoir monitoring and related algorithms are already reducing these limitations. Newer datasets—such as GRDL (Hao *et al*. 2024) and similar products that incorporate improved bathymetry reconstructions and refined simultaneous detection of open water surface area and elevation (a capability of the Surface Water and Ocean Topography Mission (Biancamaria *et al*. 2016; Fu *et al*. 2024))—represent a significant step forward. As these next-generation datasets mature, they show promise in reducing observational uncertainty and enhance the robustness of global reservoir storage evaluation. Consequently, while the limitations of current satellite-derived products must be acknowledged, they are likely to diminish in future assessments as observational technologies improve.

5. Conclusions

This study provides the first global, quantitative evaluation of reservoir storage simulations from five ISIMIP3a *global hydrological models* (GHMs), benchmarked against the satellite-derived Global Reservoir Storage (GRS) dataset to address two central questions: (1) how accurately do state-of-the-art ISIMIP GHMs reproduce global reservoir storage time series when benchmarked against satellite-derived storage observations? and (2) to what extent are GHM deficiencies attributable to parametric rigidity in generic schemes, and can these limitations be alleviated through targeted statistical bias correction or parameter tuning? Answering these questions provides the diagnostic foundation required for the improvement of storage simulations by GHMs.

The evaluation of satellite-derived storage datasets revealed that the Global Reservoir Storage (GRS) dataset demonstrated the strongest agreement with *in situ* storage observations. Although the Global Dam Storage (GDS)—our newly developed dataset—outperformed GloLakes and was a credible alternative benchmark, GRS was selected as the primary surrogate due to its superior statistical performance and a near-complete coverage of dams in the GRanD v1.3 dataset. Crucially, substituting GDS for GRS in GHM evaluation yields the same overarching patterns in model behaviour, demonstrating that our broader findings on GHMs storage characteristics are robust to the choice of GRS or GDS.

Results on model evaluation show that the current generation of ISIMIP GHMs generally struggles to simulate reservoir storage dynamics, with performance characterised by two distinct characteristics. First, the ensemble is split between dampened and exaggerated amplitudes. While MIROC-INTEG-LAND (MIL) initially performed best, its higher skill was largely a statistical artefact arising from dampened variability, which masked weak temporal correlation and significant mean bias. In contrast, the remaining GHMs—H08, WaterGAP2-2e (WGP), CWatM (CWT) and LPJml5-7-10-fire (LPJ)—systematically exhibited *over-regulation*, whereby *generic reservoir operation schemes* (GROS) classify most of the reservoirs as heavily regulated, leading to exaggerated storage amplitudes. Secondly, within the group that experience exaggerated drawdown, a clear divergence emerged between process-based and piecewise-function approaches: the GHMs employing the Hanasaki *et al*. (2006) scheme (H06)—H08, WGP and LPJ—significantly outperformed CWT, which employs the

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

LISFLOOD reservoir scheme (LIS) in temporal correlation. We demonstrate that when amplitude errors were removed through variance-matching bias correction, two things happened: firstly, the performance of all GHMs becomes generally satisfactory, and secondly, the ones with exaggerated seasonal drawdown (except H08) outperform MIL in terms of KGE, owing to their superior temporal correlation (H06-using GHMs) and mean bias performance (non-MIL GHMs except H08). We also showed that these simulation errors are not purely structural but are strongly linked to uncalibrated parameters within H06 or LIS. Sensitivity tests confirmed that tuning the *degree of regulation threshold* (DORT) effectively shifts the dam's operation regime: whereas low DORT values result in a more heavily regulated system, high DORT values progressively shift the reservoir towards inflow-tracking mode (where release becomes increasingly equal to inflow). This transition substantially improves the representation of seasonal variability and the temporal correlation. Simultaneously adjusting the *target storage level* (TSL) parameter controls the mean bias. For the LIS reservoir operating scheme, tuning the *normal storage limit parameter* (LN) set the reservoir to be operated as near empty (low LN) or near full (high LN). LN also had a significant impact on the simulated temporal correlation and variability performance of simulated storage-release. These results show that meaningful improvements can be achieved through targeted parameter calibration rather than a wholesale replacement of existing schemes.

Taken collectively, our findings signal a transition point for global reservoir modelling. The availability of accurate, long-term satellite-derived storage datasets now makes reservoir-specific parameter tuning feasible at a global scale. Moreover, it is also possible to derive upper and lower operating curves for global reservoirs using satellite-derived storage observation, offering a path to constrain parameters and storage bounds from observations. We therefore argue that future GHM development should explore hybrid architectures: robust process-based schemes provide physical representativeness, while data-driven operating curves—derived from satellite observations—act as dynamic guardrails to constrain parameters and storage simulations. This integration offers a promising pathway to reduce uncertainty in the simulation of storage and release in GHMs, enhancing global water resource assessments.

Data availability

The Global Dam Storage dataset generated in this study is available at https://doi.org/10.5281/zenodo.19465388. The *in situ* storage observations were collected from the following sources: Australia from the Australian Bureau of Meteorology (http://www.bom.gov.au/waterdata/, last access: 29 Aug 2024); India from the India Water Resources Information System (https://indiawris.gov.in/wris/#/Reservoirs, last access: 5 Feb 2025) and Tiwari and Mishra (2019); Spain from the Centro de Estudios y Experimentación de Obras Públicas (CEDEX) (https://ceh.cedex.es/anuarioaforos/afo/embalse-nombre.asp, last access: 21 Oct 2024); and the United States from the ResOpsUS dataset (Steyaert *et al*. 2022). For Canada, *in situ* reservoir water surface elevations were obtained from the Environment Canada Data Explorer (https://collaboration.cmc.ec.gc.ca/cmc/hydrometrics/www/, last access: 13 Sep 2024) and the bathymetry to convert these elevations to storage time series were obtained from the Global Reservoir Bathymetry Dataset (GRBD) (Li *et al*. 2020).

Code availability

The scripts used for computations, graphing, figure generation and statistical analyses are available at https://doi.org/10.5281/zenodo.19468213.

Author contributions

NH and EO conceptualised the study and developed the methodology. EO conducted the formal analysis, performed the H08 sensitivity tests, and wrote the original draft of the manuscript. NH supervised the research and provided project administration.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

SNG also provided project administration as the coordinator of the ISIMIP Global Water sector. NH and KO (H08), SO (LPJml5-7-10-fire), PB and LG (CWatM), YS (MIROC-INTEG-LAND), and EN (WaterGAP2-2e) performed the respective global hydrological model simulations and contributed the model outputs to the ISIMIP repository. All authors contributed to reviewing and editing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

EO and NH are financially supported by JSPS KAKENHI Grant Number 21H05002. SO's participation was supported by the Conservation International Foundation Grant Number CI-114129)

References

Biancamaria, Sylvain, Dennis P. Lettenmaier, and Tamlin M. Pavelsky. 2016. 'The SWOT Mission and Its Capabilities for Land Hydrology'. *Surveys in Geophysics* 37 (2): 307–37. https://doi.org/10.1007/s10712-015-9346-y.

Biemans, H., I. Haddeland, P. Kabat, et al. 2011. 'Impact of Reservoirs on River Discharge and Irrigation Water Supply during the 20th Century'. *Water Resources Research* 47 (3): 2009WR008929. https://doi.org/10.1029/2009WR008929.

Burek, Peter, Johan van Der Knifff, and Arie de Roo. 2013. *LISFLOOD - Distributed Water Balance and Flood Simulation Model - Revised User Manual 2013*. https://doi.org/10.2788/24719.

Burek, Peter, Yusuke Satoh, Taher Kahil, et al. 2020. 'Development of the Community Water Model (CWatM v1.04) – a High-Resolution Hydrological Model for Global and Regional Assessment of Integrated Water Resources Management'. *Geoscientific Model Development* 13 (7): 3267–98. https://doi.org/10.5194/gmd-13-3267-2020.

Busker, Tim, Ad De Roo, Emiliano Gelati, et al. 2019. 'A Global Lake and Reservoir Volume Analysis Using a Surface Water Dataset and Satellite Altimetry'. *Hydrology and Earth System Sciences* 23 (2): 669–90. https://doi.org/10.5194/hess-23-669-2019.

Chow, Ven Te, David R. Maidment, and Larry W. Mays. 1988. *Applied Hydrology*. International Edition. McGraw-Hill Series in Water Resources and Environmental Engineering. McGraw-Hill.

Döll, Petra, Howlader Mohammad Mehedi Hasan, Kerstin Schulze, et al. 2024. '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'. *Hydrology and Earth System Sciences* 28 (10): 2259–95. https://doi.org/10.5194/hess-28-2259-2024.

Döll, Petra, Frank Kaspar, and Bernhard Lehner. 2003. 'A Global Hydrological Model for Deriving Water Availability Indicators: Model Tuning and Validation'. *Journal of Hydrology* 270 (1–2): 105–34. https://doi.org/10.1016/S0022-1694(02)00283-4.

Döll, Petra, and Bernhard Lehner. 2002. 'Validation of a New Global 30-Min Drainage Direction Map'. *Journal of Hydrology* 258 (1–4): 214–31. https://doi.org/10.1016/S0022-1694(01)00565-0.

Flores, Benito E. 1986. 'A Pragmatic View of Accuracy Measurement in Forecasting'. *Omega* 14 (2): 93–98.

Frieler, Katja, Jan Volkholz, Stefan Lange, et al. 2024. 'Scenario Setup and Forcing Data for Impact Model Evaluation and Impact Attribution within the Third Round of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP3a)'. *Geoscientific Model Development* 17 (1): 1–51. https://doi.org/10.5194/gmd-17-1-2024.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

Fu, Lee-Lueng, Tamlin Pavelsky, Jean-Francois Cretaux, et al. 2024. 'The Surface Water and Ocean Topography Mission: A Breakthrough in Radar Remote Sensing of the Ocean and Land Surface Water'. *Geophysical Research Letters* 51 (4): e2023GL107652. https://doi.org/10.1029/2023GL107652.

Gao, Huilin, Charon Birkett, and Dennis P. Lettenmaier. 2012. 'Global Monitoring of Large Reservoir Storage from Satellite Remote Sensing'. *Water Resources Research* 48 (9): 2012WR012063. https://doi.org/10.1029/2012WR012063.

Grill, G., B. Lehner, M. Thieme, et al. 2019. 'Mapping the World's Free-Flowing Rivers'. *Nature* 569 (7755): 215–21. https://doi.org/10.1038/s41586-019-1111-9.

Gutenson, Joseph L., Ahmad A. Tavakoly, Mark D. Wahl, and Michael L. Follum. 2020. 'Comparison of Generalized Non-Data-Driven Lake and Reservoir Routing Models for Global-Scale Hydrologic Forecasting of Reservoir Outflow at Diurnal Time Steps'. *Hydrology and Earth System Sciences* 24 (5): 2711–29. https://doi.org/10.5194/hess-24-2711-2020.

Hanasaki, Naota, Shinjiro Kanae, and Taikan Oki. 2006. 'A Reservoir Operation Scheme for Global River Routing Models'. *Journal of Hydrology* 327 (1–2): 22–41. https://doi.org/10.1016/j.jhydrol.2005.11.011.

Hanasaki, Naota, Sayaka Yoshikawa, Yadu Pokhrel, and Shinjiro Kanae. 2018. 'A Global Hydrological Simulation to Specify the Sources of Water Used by Humans'. *Hydrology and Earth System Sciences* 22 (1): 789–817. https://doi.org/10.5194/hess-22-789-2018.

Hao, Zhen, Fang Chen, Xiaofeng Jia, et al. 2024. 'GRDL: A New Global Reservoir Area-Storage-Depth Data Set Derived Through Deep Learning-Based Bathymetry Reconstruction'. *Water Resources Research* 60 (1): e2023WR035781. https://doi.org/10.1029/2023WR035781.

Hasan, Howlader Mohammad Mehedi, Petra Döll, Seyed-Mohammad Hosseini-Moghari, Fabrice Papa, and Andreas Güntner. 2025. 'The Benefits and Trade-Offs of Multi-Variable Calibration of the WaterGAP Global Hydrological Model (WGHM) in the Ganges and Brahmaputra Basins'. *Hydrology and Earth System Sciences* 29 (2): 567–96. https://doi.org/10.5194/hess-29-567-2025.

Heinicke, Stefanie, Jan Volkholz, Jacob Schewe, et al. 2024. 'Global Hydrological Models Continue to Overestimate River Discharge'. *Environmental Research Letters* 19 (7): 074005. https://doi.org/10.1088/1748-9326/ad52b0.

Hosseini-Moghari, Seyed-Mohammad, and Petra Döll. 2025. 'The Value of Observed Reservoir Storage Anomalies for Improving the Simulation of Reservoir Dynamics in Large-Scale Hydrological Models'. *Hydrology and Earth System Sciences* 29 (17): 4073–92. https://doi.org/10.5194/hess-29-4073-2025.

Hou, Jiawei, Albert I. J. M. Van Dijk, Luigi J. Renzullo, and Pablo R. Larraondo. 2024. 'GloLakes: Water Storage Dynamics for 27 000 Lakes Globally from 1984 to Present Derived from Satellite Altimetry and Optical Imaging'. *Earth System Science Data* 16 (1): 201–18. https://doi.org/10.5194/essd-16-201-2024.

Knoben, Wouter J. M., Jim E. Freer, and Ross A. Woods. 2019. 'Technical Note: Inherent Benchmark or Not? Comparing Nash–Sutcliffe and Kling–Gupta Efficiency Scores'. *Hydrology and Earth System Sciences* 23 (10): 4323–31. https://doi.org/10.5194/hess-23-4323-2019.

Lange, Stefan, Matthias Mengel, Simon Treu, and Matthias Büchner. 2022. 'ISIMIP3a Atmospheric Climate Input Data (v1.1). ISIMIP Repository.' Version 1.1. ISIMIP Repository, October 24. https://doi.org/https://doi.org/10.48364/ISIMIP.982724.1.

Lehner, Bernhard, Penny Beames, Mark Mulligan, et al. 2024. 'The Global Dam Watch Database of River Barrier and Reservoir Information for Large-Scale Applications'. *Scientific Data* 11 (1): 1069. https://doi.org/10.1038/s41597-024-03752-9.

Lehner, Bernhard, Catherine Reidy Liermann, Carmen Revenga, et al. 2011. 'High-resolution Mapping of the World's Reservoirs and Dams for Sustainable River-flow Management'. *Frontiers in Ecology and the Environment* 9 (9): 494–502. https://doi.org/10.1890/100125.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

Li, Yao, Huilin Gao, Gang Zhao, and Kuo-Hsin Tseng. 2020. 'A High-Resolution Bathymetry Dataset for Global Reservoirs Using Multi-Source Satellite Imagery and Altimetry'. *Remote Sensing of Environment* 244 (July): 111831. https://doi.org/10.1016/j.rse.2020.111831.

Li, Yao, Gang Zhao, George H. Allen, and Huilin Gao. 2023. 'Diminishing Storage Returns of Reservoir Construction'. *Nature Communications* 14 (1): 3203. https://doi.org/10.1038/s41467-023-38843-5.

Makridakis, Spyros, Chatfield Chris, Hibon Michele, et al. 1993. 'The M2-Competition: A Real-Time Judgementally Based Forecasting Study'. *International Journal of Forecasting* 9 (1): 5–22.

Masaki, Yoshimitsu, Naota Hanasaki, Hester Biemans, et al. 2017. 'Intercomparison of Global River Discharge Simulations Focusing on Dam Operation—Multiple Models Analysis in Two Case-Study River Basins, Missouri–Mississippi and Green–Colorado'. *Environmental Research Letters* 12 (5): 055002. https://doi.org/10.1088/1748-9326/aa57a8.

Müller Schmied, Hannes, Denise Cáceres, Stephanie Eisner, et al. 2021. 'The Global Water Resources and Use Model WaterGAP v2.2d: Model Description and Evaluation'. *Geoscientific Model Development* 14 (2): 1037–79. https://doi.org/10.5194/gmd-14-1037-2021.

Müller Schmied, Hannes, Simon Newland Gosling, Marlo Garnsworthy, et al. 2025. 'Graphical Representation of Global Water Models'. *Geoscientific Model Development* 18 (April): 2409–25. https://doi.org/https://doi.org/10.5194/gmd-18-2409-2025.

Müller Schmied, Hannes, Tim Trautmann, Sebastian Ackermann, et al. 2024. 'The Global Water Resources and Use Model WaterGAP v2.2e: Description and Evaluation of Modifications and New Features'. *Geoscientific Model Development* 17 (23): 8817–52. https://doi.org/10.5194/gmd-17-8817-2024.

Nazemi, A., and H. S. Wheater. 2015a. 'On Inclusion of Water Resource Management in Earth System Models – Part 1: Problem Definition and Representation of Water Demand'. *Hydrology and Earth System Sciences* 19 (1): 33–61. https://doi.org/10.5194/hess-19-33-2015.

Nazemi, A., and H. S. Wheater. 2015b. 'On Inclusion of Water Resource Management in Earth System Models – Part 2: Representation of Water Supply and Allocation and Opportunities for Improved Modeling'. *Hydrology and Earth System Sciences* 19 (1): 63–90. https://doi.org/10.5194/hess-19-63-2015.

Nilsson, Christer, Catherine A. Reidy, Mats Dynesius, and Carmen Revenga. 2005. 'Fragmentation and Flow Regulation of the World's Large River Systems'. *Science* 308 (5720): 405–8. https://doi.org/10.1126/science.1107887.

Oberhagemann, Luke, Maik Billing, Werner Von Bloh, et al. 2025. 'Sources of Uncertainty in the SPITFIRE Global Fire Model: Development of LPJmL-SPITFIRE1.9 and Directions for Future Improvements'. *Geoscientific Model Development* 18 (6): 2021–50. https://doi.org/10.5194/gmd-18-2021-2025.

Oki, Taikan, Teruyuki Nishimura, and Paul Dirmeyer. 1999. 'Assessment of Annual Runoff from Land Surface Models Using Total Runoff Integrating Pathways (TRIP)'. *Journal of the Meteorological Society of Japan. Ser. II* 77 (1B): 235–55. https://doi.org/10.2151/jmsj1965.77.1B_235.

Oki, Taikan, and Y. C. Sud. 1998. *Design of Total Runoff Integrating Pathways (TRIP)—A Global River Channel Network*.

Otta, Kedar, Hannes Müller Schmied, Simon N. Gosling, and Naota Hanasaki. 2025. 'Towards the Use of Satellite Remote Sensing to Validate Reservoir Storage in Global Hydrological Models: Methodology and Pilot Study in the CONUS'. *Environmental Research: Water* 1 (1): 015002. https://doi.org/10.1088/3033-4942/adc47b.

Poff, N. LeRoy, J. David Allan, Mark B. Bain, et al. 1997. 'The Natural Flow Regime'. *BioScience* 47 (11): 769–84. https://doi.org/10.2307/1313099.

Sadki, Malak, Simon Munier, Aaron Boone, and Sophie Ricci. 2023. 'Implementation and Sensitivity Analysis of the Dam-Reservoir OPeration Model (DROP v1.0) over Spain'. *Geoscientific Model Development* 16 (2): 427–48. https://doi.org/10.5194/gmd-16-427-2023.

https://doi.org/10.5194/egusphere-2026-2043
Preprint. Discussion started: 18 May 2026

Steyaert, Jennie C., Laura E. Condon, Sean W.D. Turner, and Nathalie Voisin. 2022. 'ResOpsUS, a Dataset of Historical Reservoir Operations in the Contiguous United States'. *Scientific Data* 9 (1): 34. https://doi.org/10.1038/s41597-022-01134-7.

Steyaert, Jennie C., Edwin H. Sutanudjaja, Marc Bierkens, and Niko Wanders. 2025. 'Data Derived Reservoir Operations Simulated in a Global Hydrologic Model'. *Hydrology and Earth System Sciences* 29 (22): 6499–527. https://doi.org/10.5194/hess-29-6499-2025.

Sutanudjaja, Edwin H., Rens Van Beek, Niko Wanders, et al. 2018. 'PCR-GLOBWB 2: A 5 Arcmin Global Hydrological and Water Resources Model'. *Geoscientific Model Development* 11 (6): 2429–53. https://doi.org/10.5194/gmd-11-2429-2018.

Telteu, Camelia-Eliza, Hannes Müller Schmied, Wim Thiery, et al. 2021. 'Understanding Each Other's Models: An Introduction and a Standard Representation of 16 Global Water Models to Support Intercomparison, Improvement, and Communication'. *Geoscientific Model Development* 14 (6): 3843–78. https://doi.org/10.5194/gmd-14-3843-2021.

Tiwari, Amar Deep, and Vimal Mishra. 2019. 'Prediction of Reservoir Storage Anomalies in India'. *Journal of Geophysical Research: Atmospheres* 124 (7): 3822–38. https://doi.org/10.1029/2019JD030525.

Turner, Sean W. D., Jennie Clarice Steyaert, Laura Condon, and Nathalie Voisin. 2021. 'Water Storage and Release Policies for All Large Reservoirs of Conterminous United States'. *Journal of Hydrology* 603 (December): 126843. https://doi.org/10.1016/j.jhydrol.2021.126843.

United Nations Statistics Division. n.d. *Standard Country or Area Codes for Statistical Use (M49)*. United Nations. Accessed 16 March 2026. https://unstats.un.org/unsd/methodology/m49/.

Wang, Jida, Blake A. Walter, Fangfang Yao, et al. 2022. 'GeoDAR: Georeferenced Global Dams and Reservoirs Dataset for Bridging Attributes and Geolocations'. *Earth System Science Data* 14 (4): 1869–99. https://doi.org/10.5194/essd-14-1869-2022.

Wirth, Stephen Björn, Johanna Braun, Jens Heinke, et al. 2024. 'Biological Nitrogen Fixation of Natural and Agricultural Vegetation Simulated with LPJmL 5.7.9'. *Geoscientific Model Development* 17 (21): 7889–914. https://doi.org/10.5194/gmd-17-7889-2024.

Yassin, Fuad, Saman Razavi, Mohamed Elshamy, Bruce Davison, Gonzalo Sapriza-Azuri, and Howard Wheater. 2019. 'Representation and Improved Parameterization of Reservoir Operation in Hydrological and Land-Surface Models'. *Hydrology and Earth System Sciences* 23 (9): 3735–64. https://doi.org/10.5194/hess-23-3735-2019.

Yigzaw, Wondmagegn, Hong-Yi Li, Yonas Demissie, et al. 2018. 'A New Global Storage-Area-Depth Data Set for Modeling Reservoirs in Land Surface and Earth System Models'. *Water Resources Research* 54 (12). https://doi.org/10.1029/2017WR022040.

Yokohata, Tokuta, Tsuguki Kinoshita, Gen Sakurai, et al. 2020. 'MIROC-INTEG-LAND Version 1: A Global Biogeochemical Land Surface Model with Human Water Management, Crop Growth, and Land-Use Change'. *Geoscientific Model Development* 13 (10): 4713–47. https://doi.org/10.5194/gmd-13-4713-2020.

Yoshida, T., N. Hanasaki, K. Nishina, J. Boulange, M. Okada, and P. A. Troch. 2022. 'Inference of Parameters for a Global Hydrological Model: Identifiability and Predictive Uncertainties of Climate-Based Parameters'. *Water Resources Research* 58 (2): e2021WR030660. https://doi.org/10.1029/2021WR030660.

Zajac, Zuzanna, Beatriz Revilla-Romero, Peter Salamon, Peter Burek, Feyera A. Hirpa, and Hylke Beck. 2017. 'The Impact of Lake and Reservoir Parameterization on Global Streamflow Simulation'. *Journal of Hydrology* 548 (May): 552–68. https://doi.org/10.1016/j.jhydrol.2017.03.022.

Zhao, Gang, and Huilin Gao. 2018. 'Automatic Correction of Contaminated Images for Assessment of Reservoir Surface Area Dynamics'. *Geophysical Research Letters* 45 (12): 6092–99. https://doi.org/10.1029/2018GL078343.