https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

GSIM-PLUS: A Gap-Filled Global Monthly Streamflow Dataset for 1995-2015

Mingrui Chen[1,2], Bin Yong[1,2]*, Xinzhi Xu[4], Weihong Liao[3], Haichen Li[3]*, Jinxuan Zhang[4], Jinyu Xiao[4], Wei Dai[3], Jia Wang[5]

[1] State Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing 210098, China.
[2] College of Hydrology and Water Resources, Hohai University, Nanjing, 210098, China
[3] China Institute of Water Resources and Hydropower Research, Beijing, 100000, China
[4] Global Energy Interconnection Group Co. Ltd, Beijing, 100000, China
[5] School of Ocean Energy, Tianjin University of Technology, Tianjin 300384, China

Correspondence to: Bin Yong(yongbin@hhu.edu.com); Haichen Li(lihc@iwhr.com)

Abstract. Global monthly streamflow observations are fundamental for understanding changes in the water cycle, supporting large-sample hydrology, and informing water-resources assessments. However, currently available open station archives still suffer from substantial limitations in temporal continuity and spatial coverage. Here we present GSIM-PLUS, a gap-filled global monthly streamflow dataset for 1995-2015 designed to improve the completeness and reusability of global runoff records. Using the GSIM monthly archive as the basis, we identified 7,323 high-completeness anchor stations and 8,731 target stations from 30,959 gauges. Basin descriptors from five groups - climate, topography, soil, spatial location, and hydrology - were used to identify the most similar donor stations for each target site. Donor-Trend Recursive Regression (DTRR) was adopted as the default imputation method, with a guarded fallback to Model-Agnostic Meta-Learning (MAML) for a limited subset of very-low-flow stations in order to improve production stability under long recursive gaps. Multi-scenario validation shows that DTRR achieved the best overall performance under random 30 % masking (NSE = 0.864; KGE = 0.920), remained robust for 12-month continuous gaps (NSE = 0.797), and retained positive but reduced skill for very long gaps exceeding 25 months (NSE = 0.511). Comparison with independent GRDC observations at 16 co-located stations across six regions provided an external reference check, indicating that temporal co-variation was generally reproduced more reliably than exact flow magnitude, although agreement varied among stations. Under the guarded DTRR production scheme, GSIM-PLUS fills 303,271 missing monthly records for 16,054 stations, increasing the median completeness of target stations from 66.3 % to 81.0 % and that of the full dataset from 86.9 % to 95.2 %. Each released record is accompanied by quality and context metadata, including reconstruction class, gap length, fill method, and basin-context flags. GSIM-PLUS provides a more continuous and traceable global monthly streamflow resource for regional hydrological analysis, large-sample studies, model evaluation, and related monthly-scale applications. The GSIM-PLUS dataset is publicly available through Zenodo at https://doi.org/10.5281/zenodo.21425702.

Keywords: monthly streamflow; gap filling; donor-based reconstruction; global hydrological dataset; quality flags

1 Introduction

Long-term, continuous river streamflow records underpin water-cycle research, hydrological model calibration, large-sample hydrology (LSH), and regional water-resources assessment (Gupta et al., 2014; Addor et al., 2017; Alvarez-Garreton et al., 2018). Yet global observational networks have been declining steadily owing to station closures and funding cuts, creating increasingly fragmented archives that limit cross-catchment comparison and data-driven modelling (Vörösmarty et al., 2001; Hannah et al., 2011; Fekete et al., 2015; Karpatne et al.,2017).

The Global Streamflow Indices and Metadata (GSIM) project has substantially advanced global data integration. GSIM Part 1 compiled 30,959 daily streamflow time series with catchment boundaries and metadata (Do et al., 2018), and Part 2 provided multi-scale streamflow indices and homogeneity assessments (Gudmundsson et al., 2018). Nevertheless, as a directly compiled archive, the GSIM monthly series still contain pervasive gaps that reduce their reuse value for trend attribution, extreme-event analysis, and long-term modelling.

Several approaches have been pursued to mitigate streamflow data gaps. Traditional temporal interpolation (e.g., ARMA) and spatial methods (e.g., IDW) are straightforward but ignore catchment similarity (Gyau-Boakye and Schultz, 1994; Harvey et al., 2012). The Predictions in Ungauged Basins (PUB) community has demonstrated that physical similarity often outperforms spatial proximity when transferring hydrological information across basins (Oudin et al., 2008; Blöschl et al., 2013; Beck et al., 2016). Meanwhile, gridded products such as GRUN (Ghiggi et al., 2019) provide spatially continuous runoff fields but not station-scale records, and reanalysis outputs (e.g., ERA5-Land) still deviate considerably from observations in arid and small-catchment settings. The GRDC archive, though authoritative, does not perform systematic gap filling.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Consequently, a global product that fills gaps in archived monthly station records while preserving explicit quality flags and gap-context indicators remains unavailable.

GSIM-PLUS is developed to fill this gap. Building on the GSIM monthly archive for 1995-2015, this study: (1) identifies 7,323 high-completeness anchor stations and 8,731 target stations from 30,959 gauges and selects donors using 17 basin attributes spanning climate, topography, soil, spatial location, and hydrology; (2) adopts Donor-Trend Recursive Regression (DTRR) as the default imputation strategy, with a guarded fallback for near-zero-flow stations, and validates performance under random, continuous, mixed, and very-long-gap scenarios; and (3) releases a gap-filled monthly streamflow dataset equipped with a Q0-Q3 quality-flag system and station-level gap-context fields, providing a more continuous and reusable global observational resource.

GSIM-PLUS should be understood as an independent derivative data product based on the publicly available GSIM monthly archive, rather than as an official update of GSIM itself. Its purpose is to improve the temporal continuity and practical usability of monthly station records while preserving traceability to the underlying GSIM observations and metadata.

2 Data

GSIM-PLUS was constructed over a common analysis window of January 1995 to December 2015 (252 months). This period was chosen because it offers the strongest overlap among national monitoring networks while still being long enough to capture interannual hydrological variability.

2.1 GSIM monthly streamflow records

Monthly mean streamflow records (m^3 s^{-1}) from the GSIM database (Do et al., 2018; Gudmundsson et al., 2018) serve as the base observational source. Of the 30,959 stations in GSIM, those with records falling within the common window were extracted and placed on a fixed monthly timeline. Basic quality checks were applied before product generation, including removal of negative values, time-axis continuity verification, and flagging of abrupt anomalies. Original observations, missing-value states, and station metadata were retained together as the consistent basis for all subsequent processing steps.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

2.2 Basin attributes used for donor matching

To identify hydrologically comparable donor stations, 17 basin attributes were compiled from multiple global datasets and organized into five groups (Tab.1). These attributes characterize the background hydrological similarity among stations and are used exclusively for donor selection, not as direct inputs to the imputation model.

Table 1. Basin attributes and data sources used in GSIM-PLUS.

Group	Variables	Source	Reference	Data access information
Climate (7)	Tp_mean, Tp_std, Tp_cv, t2m_mean, t2m_std, E_mean, E_std	ERA5-Land	Muñoz-Sabater et al. (2021)	cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels
Topography (2)	Altitude; Slope	SRTM; HydroSHEDS	Farr et al. (2007)	earthdata.nasa.gov/data/instruments/srtm hydrosheds.org/products/hydrorivers; hydrosheds.org/products/hydrobasins
Soil (3)	Sand, Silt and Clay fractions	GSIM catchment characteristics	Do et al. (2018,data set)	doi.pangaea.de/10.1594/PANGAEA.887477
Spatial (2)	Latitude; Longitude	Station coordinates	Do et al. (2018,data set)	doi.pangaea.de/10.1594/PANGAEA.887477
Hydrology (3)	Mean flow; Upstream area; Local area	HydroSHEDS	Lehner and Grill (2013)	hydrosheds.org/products/hydrorivers; hydrosheds.org/products/hydrobasins

Catchment boundaries from HydroSHEDS were used to aggregate gridded climate, topographic, and soil fields to the station scale. The Köppen–Geiger climate classification was adopted for regional performance stratification in subsequent analyses. All inputs were harmonized in terms of naming conventions, units, and missing-value encoding before entering the production workflow.

2.3 Independent GRDC reference data

Monthly observations from the Global Runoff Data Centre (GRDC; portal.grdc.bafg.de) were matched to 16 co-located GSIM stations spanning six regions and used as an independent reference (Sect. 4.4). The GRDC values were not used in donor selection, model fitting, or reconstruction. Because the corresponding GSIM stations remained part of the released station set, this analysis is an independent-observation comparison rather than a station-wise holdout experiment.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

3 Reconstruction framework

3.1 Station screening and reconstruction design

The GSIM-PLUS production workflow was developed for monthly streamflow records over 1995-2015. We began with 30,959 GSIM stations and classified them into anchor, target, and excluded groups according to record completeness within the study period. Stations with at least 90 % monthly availability were treated as anchor stations and used as donor candidates in the reconstruction workflow. Stations with completeness between 30 % and 90 % were treated as target stations and submitted to gap filling. Stations below the lower threshold, or otherwise not suitable for reconstruction, were excluded from product generation. Under these rules, the final station inventory consists of 7,323 anchor stations, 8,731 target stations, and 14,905 excluded stations.

Because method evaluation requires a sufficient number of observed months after masking, a stricter evaluable subset was further defined within the target group. Specifically, target stations with at least 120 observed months during 1995-2015 were retained for validation experiments, yielding 6,978 evaluable target stations. This distinction is important throughout the manuscript: the full target set is used for product generation, whereas the evaluable target subset is used to assess reconstruction skill under controlled missing-data experiments. In this way, the validation design remains tied to the released product while avoiding performance estimates based on stations with too little remaining information.

The overall design therefore separates the workflow into three layers: station screening, donor matching and reconstruction, and product-level quality annotation. This structure keeps the production logic transparent and ensures that each reconstructed value can be traced back to both its donor context and its gap characteristics.

3.2 Donor matching and production reconstruction workflow

For each target station, donor selection was based on 17 basin attributes grouped into hydrology, climate, topography, spatial location, and soil properties. To define the similarity metric, we evaluated several plausible weighting configurations for both the five feature groups and the relative importance of variables within each group. Because downstream validation results were highly stable across these alternative configurations, we adopted a fixed-prior weighting scheme derived from an analytic hierarchy process (AHP) informed by hydrological domain knowledge. The released data package includes the final feature-weight tables, donor-matching metadata, and the parameter-sensitivity outputs used to document the robustness of the weighting design (see Data availability). In the final scheme, group-level weights of 0.40, 0.25, 0.15, 0.10, and 0.10 were assigned to hydrology, climate, topography, spatial location, and soil properties, respectively. Within each group, fixed prior weights were further assigned to individual variables to reflect their relative hydrological importance. Specifically, the hydrology group used prior weights of 0.50, 0.35, and 0.15 for mean flow, upstream area, and local basin area; the climate group assigned larger prior weights to precipitation statistics than to evaporation variability; the topography group slightly favored slope over elevation; the spatial group treated latitude and longitude equally; and the soil group assigned nearly equal weights to sand, silt, and clay fractions. No data-driven fitting or post hoc weight tuning was applied. A lightweight sensitivity analysis using the baseline and four alternative weighting configurations showed only marginal changes in downstream DTRR skill across reasonable perturbations of the prior weights; the complete scheme definitions and results are included in the released package.

Before similarity calculation, hydrological variables were transformed using log(1+x) to reduce scale imbalance. Missing feature values were filled using the median of the anchor-station feature distribution, and all attributes were standardized using anchor-based z-score scaling. Similarity between a target station and a candidate donor was then defined as a monotonic transformation of weighted Euclidean distance in the standardized feature space. To improve hydroclimatic consistency, donor search preferentially retained anchors from the same continent and Koppen-Geiger class as the target station. If fewer than five donors were available under this constraint, the search was progressively relaxed to anchors from the same continent and Koppen code prefix, then the same continent only, then the same Koppen major class, then the same Koppen code prefix, and finally the full anchor pool. The five most similar anchors were retained as the donor pool for each target station, yielding 43,655 donor-target pairs with a mean similarity of 0.909.

Gap filling was then carried out using Donor-Trend Recursive Regression (DTRR) as the default production method. DTRR is implemented as a station-specific ridge-regression model fitted on the standardized observed record of each target station. For each missing month, the model combines seasonal terms, target persistence, and donor-based information derived from the three highest-similarity donors within the matched donor pool:

$$\hat{z}_t = \beta_0 + \boldsymbol{\beta}^{\top}\mathbf{x}_t,$$

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

where $\hat{z}_t$ is the standardized reconstructed flow at month t , and $\mathbf{x}_t$ includes month-of-year harmonics, the lag-1 target flow, the current flow and month-to-month change of the top three donors, a similarity-weighted donor mean, a similarity-weighted donor change, and donor coverage at that time step. Each prediction is transformed back to the original flow scale and constrained to be nonnegative. The constrained value is then restandardized, written back to the target series, and used as the lag-1 input for the next missing month, making the reconstruction explicitly recursive.

To prevent numerical divergence in very-low-flow settings, a guarded fallback was introduced in the production workflow. For stations with median flow below 0.02 $m^3 s^{-1}$, the released product uses Model-Agnostic Meta-Learning (MAML) instead of DTRR. MAML is a neural regression model with seasonal harmonics and lag-1 target flow as inputs. Its first-order meta-initialization is learned exclusively from anchor stations using balanced simulated random 30 %, 3-, 6-, 12-, and 25–48-month gaps. For each meta-training task, station-level standardization is estimated from support observations only, and hidden query months are predicted recursively. At each target station, the learned initialization is adapted using matched-anchor observations and the available target record before sequential gap reconstruction. All reconstructed flows are constrained to be nonnegative before recursive feedback. MAML therefore also uses recursive lag-1 feedback; its safeguard role is based on empirical numerical stability and its consistently highest NSE among the non-DTRR methods across the validation scenarios in Sect. 4.3, rather than on removal of recursion. The 0.02 $m^3 s^{-1}$ threshold remains a pragmatic production safeguard rather than a theoretically unique cutoff.

3.3 Quality-control design and product metadata

The GSIM-PLUS quality-control design follows a two-layer logic. The first layer records reconstruction provenance and gap context directly in the main time-series product. Each record includes “observed_streamflow”, “final_streamflow”, “ segment_length”, “fill_method” and “quality_flag”. Here, “quality_flag” distinguishes observed values from reconstructed values and links each reconstructed point to the length of the missing segment in which it occurs: “Q0” denotes original observations, “Q1” denotes reconstructed values within short gaps of up to 3 months, “Q2” denotes reconstructed values within intermediate gaps of 4-24 months, and “Q3” denotes reconstructed values within gaps longer than 24 months. If no stable estimate can be produced, the framework reserves “Q4” for unresolved values.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

The “fill_method” field records whether a value was preserved as `OBSERVED` or reconstructed using “DTRR”, “MAML”, or another designated method in comparison experiments.

The second layer records contextual risk. In the current release, this layer is also written directly into the product tables through “kg_major”, “kg_code”, “arid_flag”, “low_flow_flag”, “context_risk_flag”, and “guard_applied”. These fields identify the hydroclimatic setting of each station, indicate whether the station belongs to the arid climate class, mark very-low-flow stations with median flow below 0.02 m^3 s^{-1}, summarize the combined contextual risk status, and record whether the low-flow safeguard was activated in production. In this way, the released archive does not rely on a single scalar uncertainty value, but instead provides users with a compact set of provenance and context indicators that can be used together for downstream filtering and interpretation.

This two-layer design is important because monthly reconstruction uncertainty is not controlled by one factor alone. Gap length is a first-order indicator of reconstruction difficulty, but it does not fully capture hydroclimatic regime or low-flow instability. In GSIM-PLUS, the “Q0-Q3” system therefore serves as the primary provenance tier, while the contextual fields provide a complementary risk tier for interpretation and screening. Users can first select records by reconstruction class and then refine their selection according to basin setting and low-flow context most relevant to their application.

4 Validation of the production workflow

4.1 Dataset overview

The final GSIM-PLUS product contains 16,054 stations, including 7,323 anchor stations and 8,731 target stations, and spans all inhabited continents. Under the guarded DTRR production workflow, the dataset fills 303,271 missing monthly records, of which 277,729 belong to target stations. Spatial density is highest in Europe and eastern North America and lowest in central Africa and central Asia (Fig. 1), reflecting the underlying distribution of the GSIM archive. Detailed completeness changes and quality-flag distributions are presented in Sect.5.2.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

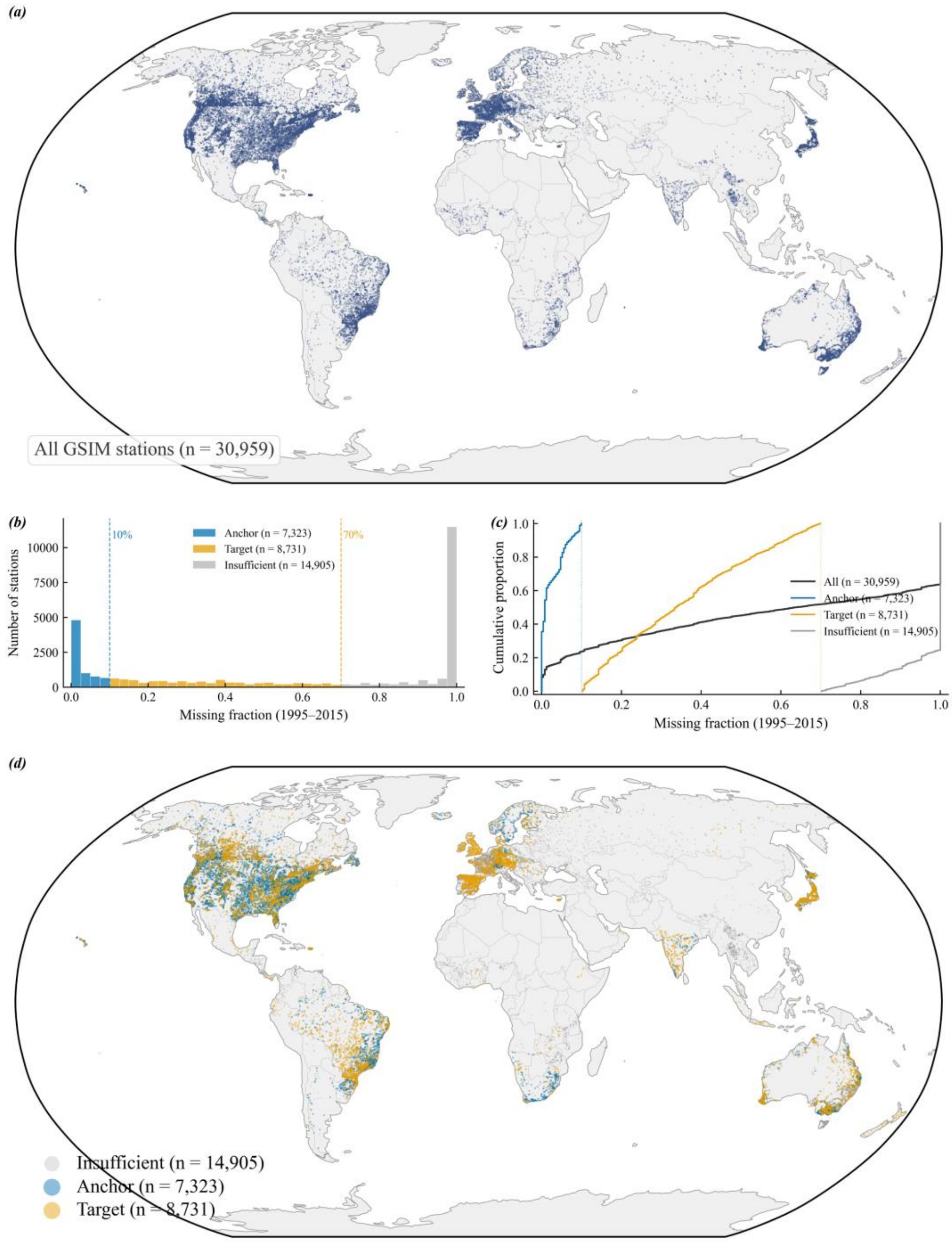


Figure 1. Global station distribution and station screening. **(a)** Global distribution of all GSIM stations. **(b)** Number of stations by missing-rate class for anchor, target, and insufficient-observation groups. **(c)** Cumulative proportion of stations as a function of missing rate for the three station groups. **(d)** Global distribution of the station classes defined in this study.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

4.2 Data gaps and evaluation overview

The validation framework was designed to reflect the main types of missingness observed in global monthly streamflow archives rather than to function as a generic method-comparison benchmark. In practice, GSIM target stations contain a mixture of isolated missing months, short continuous outages, medium-length interruptions, and occasional very long data gaps. These different gap structures imply different reconstruction risks and directly motivate the quality classes used in GSIM-PLUS. Accordingly, the production workflow was evaluated under four complementary validation settings: random masking, fixed-length continuous gaps, mixed hybrid gaps, and super-long gaps.

The random 30 % masking experiment represents scattered missingness under otherwise well-observed conditions. The continuous-gap experiments with 3, 6, and 12 missing months represent short outages, seasonal-scale interruptions, and year-scale breaks, respectively. The hybrid experiments combine multiple gap length and are intended to mimic the mixed missingness patterns commonly found in operational archives. The operating conditions are set as shown in Table 2. Finally, the super-long-gap experiment targets the most difficult reconstruction regime, with missing segments longer than 25 months. Together, these scenarios define the practical quality boundary of the product: they test not only whether DTRR performs well on average, but also under what kinds of missing-data structures reconstruction remains reliable enough for inclusion in the released dataset.

Table 2. The operating conditions of hybrid experiments.

Category	1 month	3 months	6 months	12 months	24 months	25+ months
H1 (%)	35	30	20	10	5	0
H2 (%)	20	20	20	15	15	10
H3 (%)	10	10	15	20	20	25

Eight candidate methods were evaluated in the internal validation experiments: two simple baselines (Seasonal Mean and IDW), four general machine-learning regressors (Random Forest, Linear Regression, KNN, and LSTM), a gap-aware neural meta-learning model (MAML), and the production method (DTRR). These comparisons are used to justify the production choice and identify the conditions under which reconstruction uncertainty begins to rise, rather than to rank methods as an end in itself.

4.3 Overall validation summary

Table 3 summarizes the main internal-validation results for the production workflow. Across all evaluated missing-data structures, DTRR provided the highest overall skill and the most stable behavior as gap

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

difficulty increased. Under random 30 % masking, DTRR achieved NSE = 0.864 and KGE = 0.920, exceeding the best non-DTRR alternative. In the continuous-gap experiments, DTRR remained the top method for all three scenarios, with NSE/KGE values of 0.879/0.865 for 3-month gaps, 0.930/0.931 for 6-month gaps, and 0.797/0.869 for 12-month gaps. The 6-month scenario yielded the highest pooled skill, indicating that performance is not strictly monotonic with gap length under this validation design; because each scenario uses independently sampled gap blocks and pooled evaluation points, the reported statistics reflect scenario difficulty rather than a strict paired comparison. Even so, the 12-month experiment shows a clear decline relative to the shorter-gap scenarios and provides a useful reference for the upper end of moderate-confidence reconstruction.

The hybrid scenarios provide a more realistic representation of archive-like missingness. In the short-gap-dominant H1 scenario, DTRR achieved NSE = 0.774 and KGE = 0.861. In the balanced H2 scenario, after excluding one unstable near-zero-flow outlier station (AU_0002125), DTRR reached NSE = 0.809 and KGE = 0.881. In the long-gap-dominant H3 scenario, again excluding AU_0002125, DTRR retained useful skill with NSE = 0.729 and KGE = 0.824. These results indicate that DTRR is effective not only for scattered or short gaps, but also for the mixed gap structures that dominate real-world station archives.

The super-long-gap experiment defines the most conservative quality boundary of the released product. After excluding three unstable outlier stations (AU_0002125, AU_0000492, and BR_0000375), DTRR achieved NSE = 0.511 and KGE = 0.715 over 391,519 common evaluation points. Although all methods degraded in this regime, DTRR retained the highest skill and MAML was the strongest non-DTRR alternative. The lower performance under these extreme conditions supports use of long-gap reconstruction in GSIM-PLUS while justifying stronger caution for the longest segments in downstream applications.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Table 3. Overall validation performance of DTRR and the strongest non-DTRR comparator.

Validation setting	DTRR NSE	DTRR KGE	Strongest non-DTRR method	Interpretation for GSIM-PLUS
Random 30 % masking	0.864	0.920	MAML	Supports DTRR as the default production method under general missingness
Continuous 3 months	0.879	0.865	MAML	High-confidence reconstruction for short gaps
Continuous 6 months	0.930	0.931	MAML	Very strong pooled skill; still within the high-confidence range
Continuous 12 months	0.797	0.869	MAML	Clear degradation begins, but performance remains strong
Hybrid H1	0.774	0.861	MAML	Robust under realistic mixed missingness dominated by short gaps
Hybrid H2	0.809	0.881	MAML	Strong performance for broadly representative archive conditions
Hybrid H3	0.729	0.824	MAML	Skill remains useful but uncertainty increases in long-gap archives
Super-long gaps (>25 months)	0.511	0.715	MAML	Defines the lower-confidence boundary of the released product

Taken together, these experiments support two product-level conclusions. First, DTRR is the appropriate default production method for GSIM-PLUS because it remains consistently superior across the major missing-data structures considered here, while MAML provides a numerically stable safeguard for the predefined very-low-flow subset. Second, reconstruction uncertainty is governed primarily by gap structure, with a marked transition from high-confidence short and medium gaps to more uncertain very long gaps. This pattern directly motivated the gap-length-based quality classes described in Sect. 3.3 and informs the usage recommendations presented in Sect. 5.2.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

4.4 Independent-reference comparison against GRDC observations

To independently evaluate the reconstructed records, GSIM-PLUS estimates were compared with GRDC monthly observations at 16 co-located stations across six continents. The analysis was restricted to months originally missing from GSIM but observed by GRDC. The GRDC values were used only for evaluation and were not involved in donor selection, model fitting, or reconstruction.

Overall reconstruction accuracy was assessed primarily using NSE, while R was used as a complementary measure of temporal co-variation. 11 of the 16 stations had positive NSE values (Table 4; Fig. 2). Strong overall agreement was obtained at BR_0000276 ($n = 24$, NSE = 0.93, $R = 0.97$, $p < 0.001$) and US_0002812 ($n = 10$, NSE = 0.79, $R = 0.92$, $p < 0.001$). FR_0001112 also produced high values ($n = 6$, NSE = 0.92, $R = 0.97$, $p = 0.0017$), although its relatively small sample requires cautious interpretation. These examples show that GSIM-PLUS can reproduce both flow magnitude and temporal variation when suitable donor information and representative target observations are available.

Twelve stations showed statistically significant positive correlations, but high R did not always coincide with high NSE. For example, US_0004183 had $R = 0.98$ but NSE = 0.50, while ZA_0000030 had $R = 0.91$ but NSE = 0.08. These results indicate that temporal variations were reproduced more accurately than flow levels, variances, or peak magnitudes at some stations, as also shown by the scatter plots and Taylor diagram (Figs. 3-4). For US_0005774, however, the high correlation ($R = 0.99$) was based on only three paired observations and was not statistically significant ($p = 0.0897$); this station is therefore not used as evidence of strong agreement. Table 4 reports the sample size and two-sided p value associated with R for all 16 stations, and correlations based on very small samples should be interpreted cautiously.

The most extreme negative NSE values occurred at AU_0000817 and ZA_0000051, for which the mean GRDC flows during the evaluated months were only 0.0034 and 0.0242 $m^3 s^{-1}$, respectively. The low variance of these near-zero reference values increased the sensitivity of NSE to reconstruction errors, consistent with the low-flow reconstruction difficulty identified in Sect. 5.1.2. However, their GSIM source-record medians exceeded the station-level 0.02 $m^3 s^{-1}$ safeguard threshold. These cases show that episodic near-zero-flow periods may occur at stations not classified as persistently low flow and that differences between the flow distributions represented by the available GSIM observations and the GRDC evaluation months can further affect magnitude agreement. Overall, the independent GRDC comparison supports the reconstruction

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

capability of GSIM-PLUS at many stations, while identifying small samples and episodic near-zero-flow conditions as important limitations.

295 **Table 4. Agreement between GSIM-PLUS reconstructions and independent GRDC observations at 16 co-located stations.**

Station	Region	Evaluated infilled months (n)	Mean flow	Fill NSE	Fill R	Fill RMSE	Fill MAE	P value
AU_0000756	Australia	7	0.85	0.67	0.8864	0.25	0.20	0.0078
AU_0000768	Australia	17	0.94	0.09	0.3878	1.49	0.60	0.1241
AU_0000817	Australia	21	<0.01	-30.64	0.1713	0.09	0.06	0.4578
BR_0000276	South America	24	6380.9	0.93	0.9739	1127.58	805.63	<0.001
BR_0000649	South America	20	11.68	-0.35	0.1497	10.96	8.58	0.5288
BR_0000662	South America	12	53.39	0.48	0.9477	72.51	30.69	<0.001
FR_0001112	Europe	6	1.25	0.92	0.9666	0.44	0.39	0.0017
IN_0000071	Asia	7	418.3	-0.05	0.9519	673.96	633.43	<0.001
JP_0000268	Asia	60	316.0	0.54	0.7562	113.97	77.10	<0.001
JP_0000805	Asia	36	107.9	0.10	0.4387	55.90	43.23	0.0074
US_0002812	North America	10	85.50	0.79	0.9195	42.25	26.85	<0.001
US_0004183	North America	14	45.04	0.50	0.9758	15.79	10.98	<0.001
US_0005774	North America	3	1757.8	0.83	0.9901	727.64	594.52	0.0897
ZA_0000030	Africa	24	40.75	0.08	0.9122	26.80	17.91	<0.001
ZA_0000048	Africa	11	0.74	-0.57	0.8021	1.03	0.83	0.003
ZA_0000051	Africa	115	0.02	-5243.36	0.2187	12.60	10.02	0.0189

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

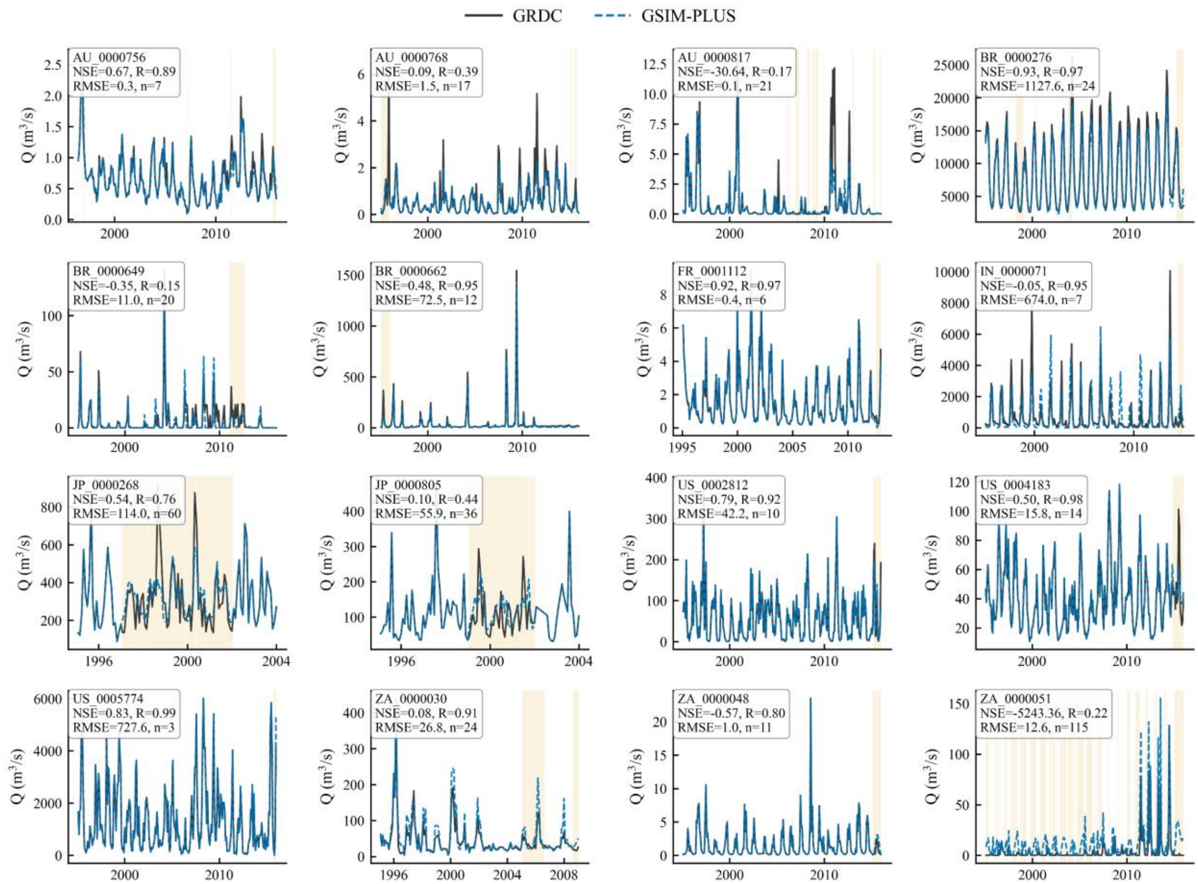
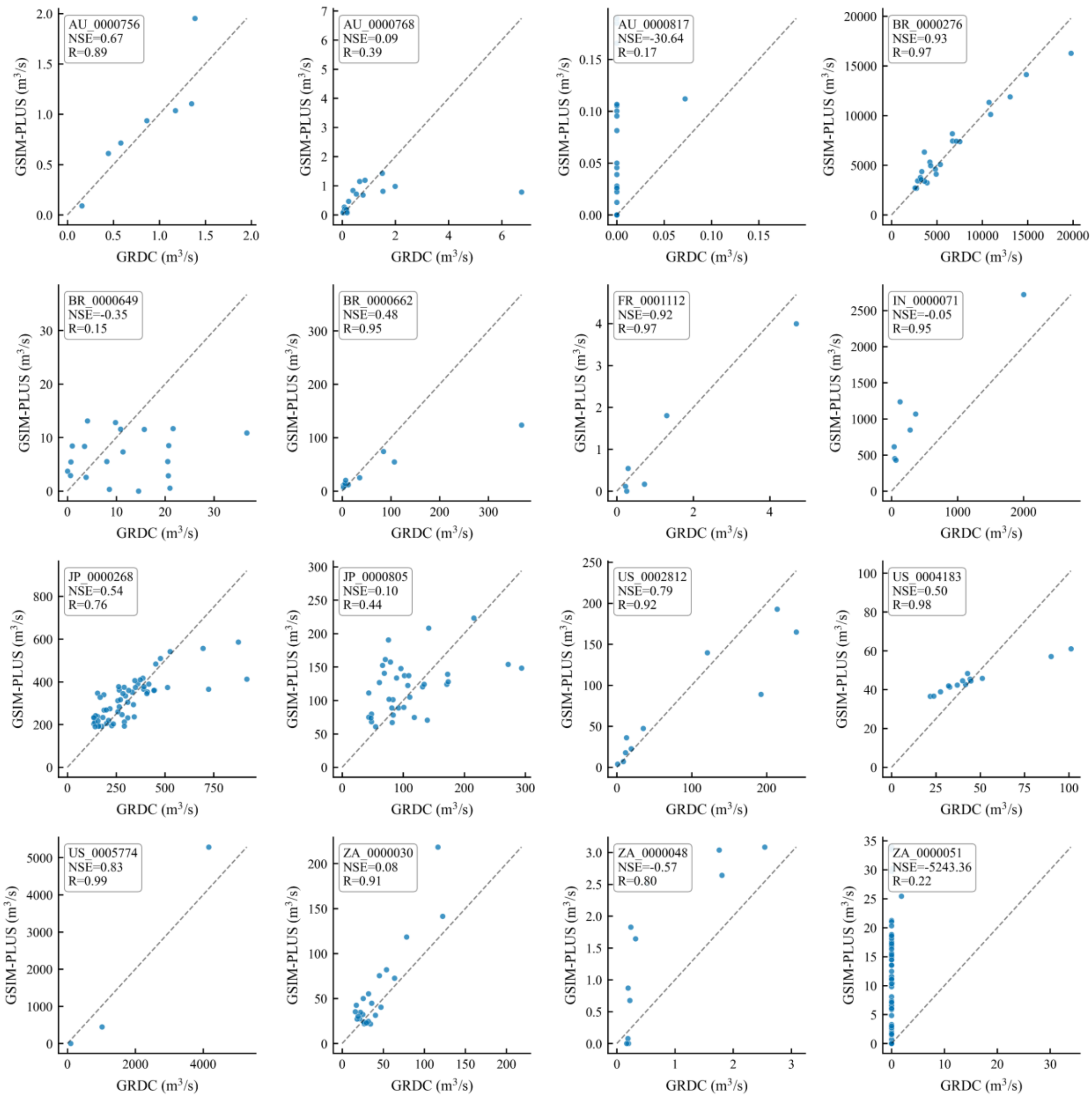


Figure 2. Time-series comparison between GSIM-PLUS reconstructions and independent GRDC observations.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions



300 **Figure 3. Scatter-plot comparison between GSIM-PLUS reconstructions and independent GRDC observations.**

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

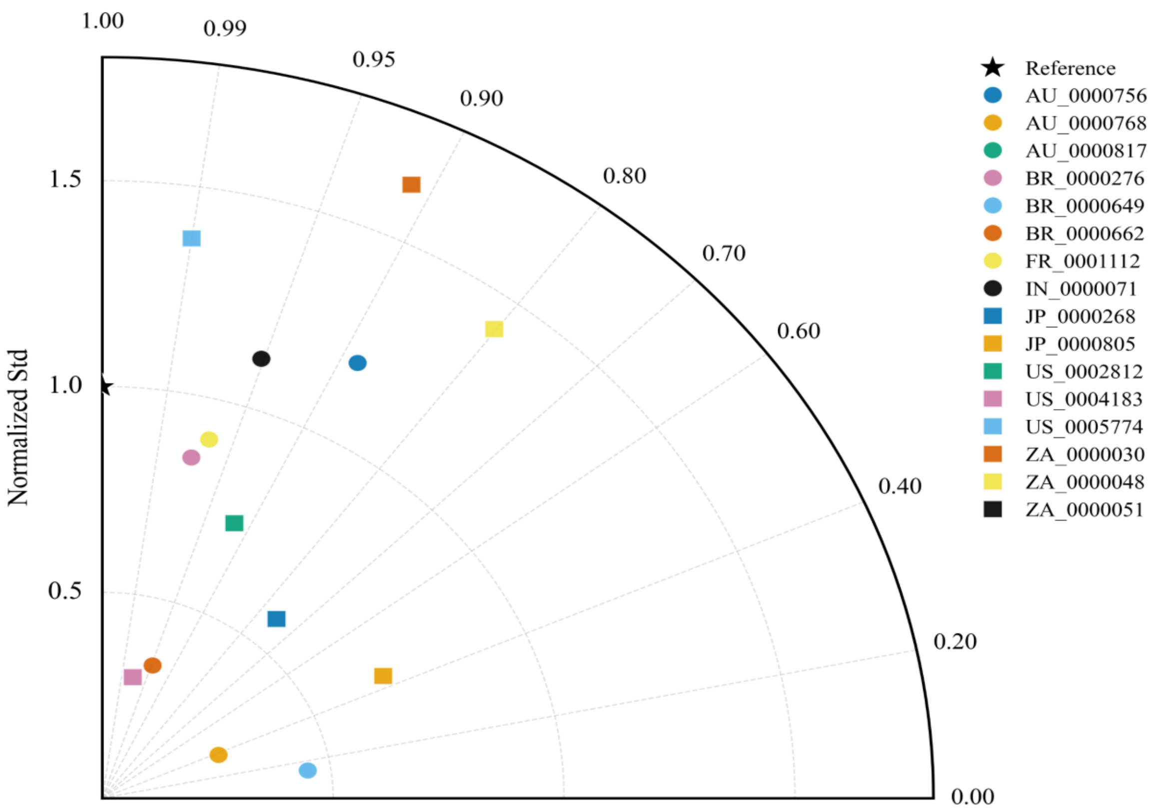


Figure 4. Taylor-diagram summary of agreement with independent GRDC observations.

5 GSIM-PLUS data records and product description

5.1 Uncertainty and quality-boundary analysis

The uncertainty structure of GSIM-PLUS is best interpreted in two linked layers. The first layer is reconstruction provenance, represented by the Q0-Q3 quality classes and associated gap metadata. The second layer is contextual basin risk, in which hydroclimatic regime and very low-flow behavior can further modify the reliability of reconstructed values within the same Q class. The analyses below therefore focus on the two clearest contextual uncertainty signals in the current product: hydroclimatic differences across climate zones and instability in very low-flow basins under long gaps.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

5.1.1 Hydroclimatic uncertainty across climate zones

Hydroclimatic differences provide the first major contextual uncertainty signal in GSIM-PLUS. The climate-zone analysis shows that DTRR remains the best-performing method across all five major Koppen-Geiger classes, but the level and spread of reconstruction skill are not spatially uniform (Fig. 5(a)). For NSE, the median performance is highest in polar and tropical basins (0.718 and 0.653), remains moderate in continental and temperate basins (0.581 and 0.496), and is lowest in arid basins (0.307). This pattern indicates that reconstruction is generally more reliable in humid and cold environments than in water-limited basins, where donor mismatch and intermittent flow behavior are more difficult to constrain.

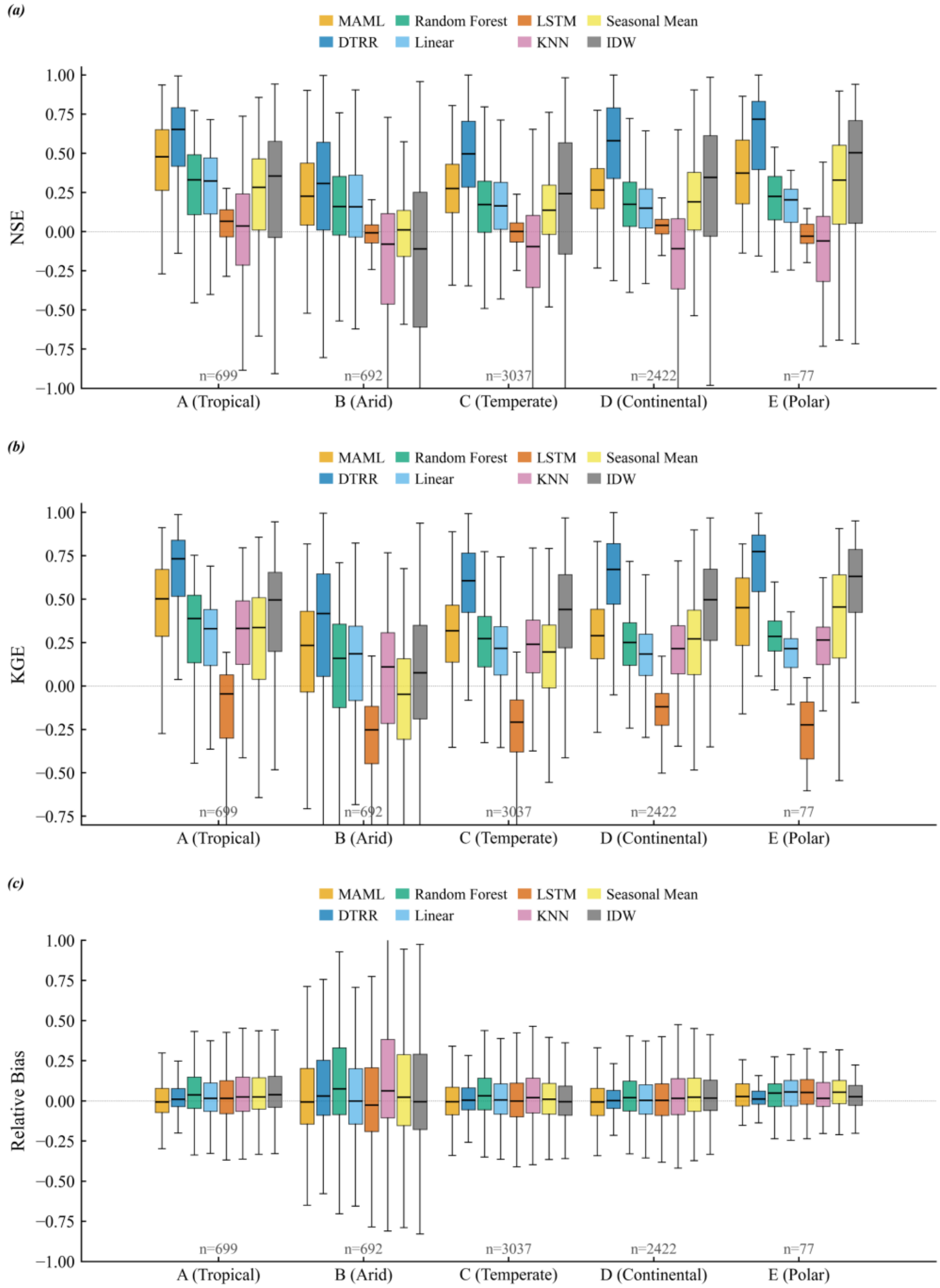
The KGE results show a similar hydroclimatic gradient (Fig. 5(b)). Median KGE reaches 0.775 in polar basins and 0.733 in tropical basins, compared with 0.672 in continental basins, 0.607 in temperate basins, and only 0.418 in arid basins. Because KGE reflects correlation, variability, and bias jointly, this pattern suggests that DTRR preserves temporal structure and flow variability more consistently in humid, tropical, cold, and polar settings than in arid regions.

Bias provides an additional perspective on uncertainty boundaries (Fig. 5(c)). Although median Bias remains relatively small compared with the full range of observed flows, the spread of Bias is much larger in arid and low-flow settings than in humid regions, indicating that systematic over- or underestimation is more likely in dry-climate basins. This larger dispersion is consistent with the stronger intermittency and donor limitation of arid rivers and suggests that these basins should be treated as the clearest hydroclimatic high-risk regime in the current release.

Taken together, Fig. 5 should be interpreted not only as a climate-zone performance comparison but also as a product-level uncertainty map. The key implication is that GSIM-PLUS is broadly usable across all major climate zones, but Q2-Q3 values in arid basins should be used more cautiously than records of the same class in humid temperate, continental, tropical, or polar basins.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions



335 **Figure 5.** Performance across different climate zones. Reconstruction performance across five major Koppen-Geiger climate zones for eight gap-filling methods, shown for (a) NSE, (b) KGE, and (c) relative Bias. DTRR performs best overall, but uncertainty is climate-dependent: arid basins show the lowest reconstruction skill and the widest bias dispersion, whereas humid and cold basins are generally more stable.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

5.1.2 Low-flow and long-gap instability boundary

Low-flow and intermittent-flow basins represent a second major uncertainty boundary in GSIM-PLUS. To quantify this effect, we examined station-level DTRR performance in the 12-month continuous-gap experiment as a function of station median flow (Fig. 6). For stations with median flow < 0.005 $m^3 s^{-1}$, median station-level NSE was -0.14 and 47.1 % of stations had NSE < 0. In the 0.005-0.01 and 0.01-0.02 $m^3 s^{-1}$ ranges, median NSE remained below zero (-0.09 and -0.04), and the fractions with NSE < 0 were 52.6 % and 50.0 %, respectively. By contrast, stations with median flow > 0.05 $m^3 s^{-1}$ had median NSE = 0.49 and an NSE < 0 fraction of 24.1 %. These results identify a clear transition in reliability around 0.02 $m^3 s^{-1}$.

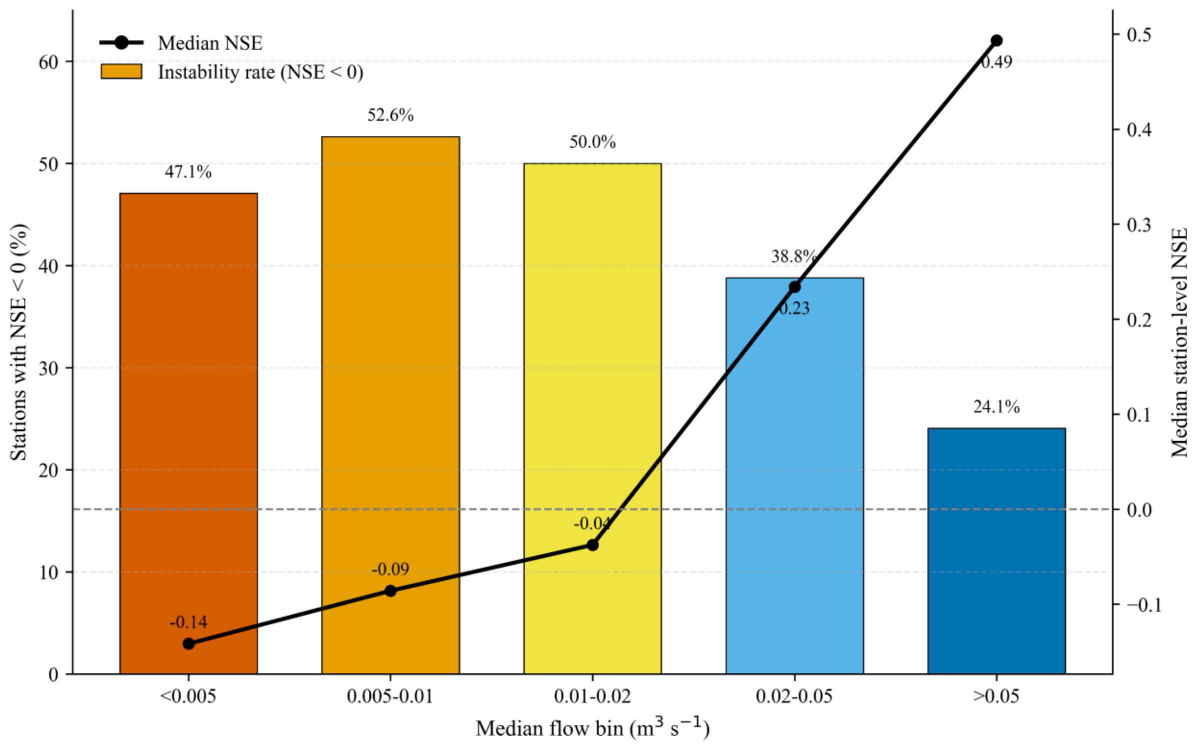


Figure 6. DTRR performance in the 12-month continuous-gap experiment as a function of station median flow

These results indicate that the transition from relatively stable to systematically fragile reconstruction occurs at approximately 0.02 $m^3 s^{-1}$. We therefore adopt median flow < 0.02 $m^3 s^{-1}$ as a low-flow safeguard threshold in the production workflow. This threshold should be interpreted as an empirical production boundary rather than a universal hydrological threshold. Its role is to identify stations for which recursive donor-trend

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

prediction is especially vulnerable to instability under long-gap conditions and to mark these records with an additional low-flow context in the released product.

The low-flow safeguard is complementary to, rather than redundant with, the Q0-Q3 quality classes. The Q classes encode reconstruction provenance and gap-length context, whereas the low-flow threshold identifies an additional basin-level instability mode that can occur even within the same Q class. In the current target-station inventory, 690 stations (7.9 % of all target stations) fall below this threshold, and 676 stations trigger the guarded fallback during production. The low-flow flag therefore applies to a limited but non-negligible subset of the archive and provides a practical second-layer warning for users working in intermittent and water-limited basins.

5.1.3 Layered interpretation of product uncertainty

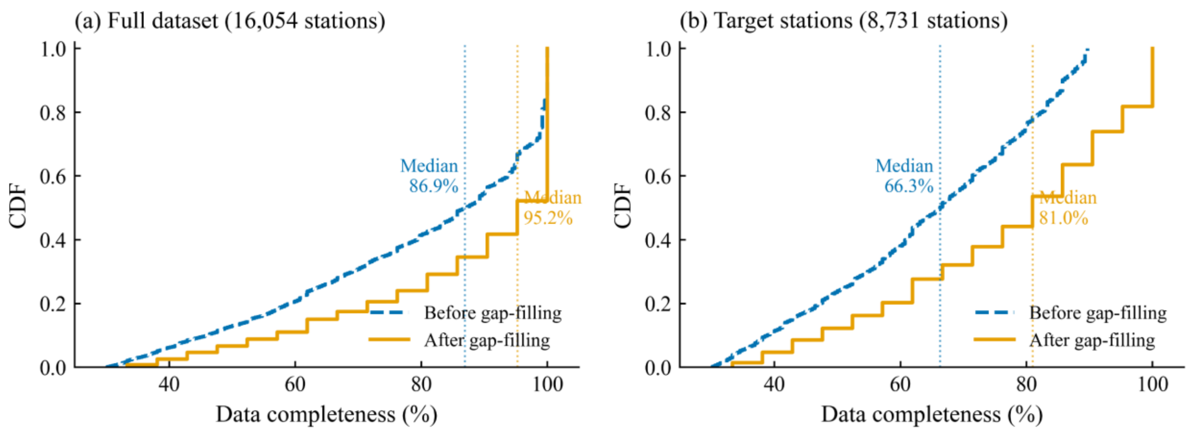
Taken together, these results suggest that GSIM-PLUS uncertainty should be interpreted in two linked layers. The first layer is defined by reconstruction provenance and gap length through the Q0-Q3 quality classes. The second layer is defined by basin context, with arid basins and very low-flow stations carrying additional risk even within the same Q class. This layered interpretation underlies the product metadata described in Sect. 3.3 and provides the basis for the practical usage recommendations discussed below.

5.2 GSIM-PLUS product overview

GSIM-PLUS provides a gap-filled global monthly streamflow archive for 1995-2015 covering 16,054 stations, including 7,323 anchor stations and 8,731 target stations. Under the guarded DTRR production workflow, a total of 303,271 monthly missing values are reconstructed. The released archive remains strongly observation-dominated, but its continuity is substantially improved in a transparent and quality-traceable manner.

The completeness statistics(Fig.7) illustrate the practical effect of the release. Across the full archive, the median completeness increases from 86.9 % before filling to 95.2 % after filling. For target stations alone, the median completeness increases from 66.3 % to 81.0 %. A total of 7,674 stations (47.8 % of the full archive) reach 100 % completeness after filling, including 1,590 target stations (18.2 % of the target subset). These gains substantially expand the number of monthly station records that can be used for climatological comparison, long-term hydrological assessment, and large-sample analysis.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026



380

Figure 7. Comparison of record completeness before and after gap filling. Empirical cumulative distribution functions (CDFs) of station-level data completeness for (a) the full dataset (16,054 stations) and (b) the target-station subset (8,731 stations), comparing conditions before and after gap-filling.

The quality composition(Fig.8) of the released product clarifies where the added continuity comes from. Across the full archive, Q0 records account for 91.4 % of all monthly values, while Q1, Q2, and Q3 account for 1.3 %, 3.7 %, and 3.7 %, respectively. Within the target-station subset, reconstruction naturally occupies a larger share: Q0 = 83.7 %, Q1 = 2.0 %, Q2 = 6.7 %, and Q3 = 7.6 %. This shift is expected because the target stations are precisely those with more fragmented original records. Even so, the product remains dominated by original observations, with reconstructed values concentrated mainly in the medium- and long-gap portions of the archive.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

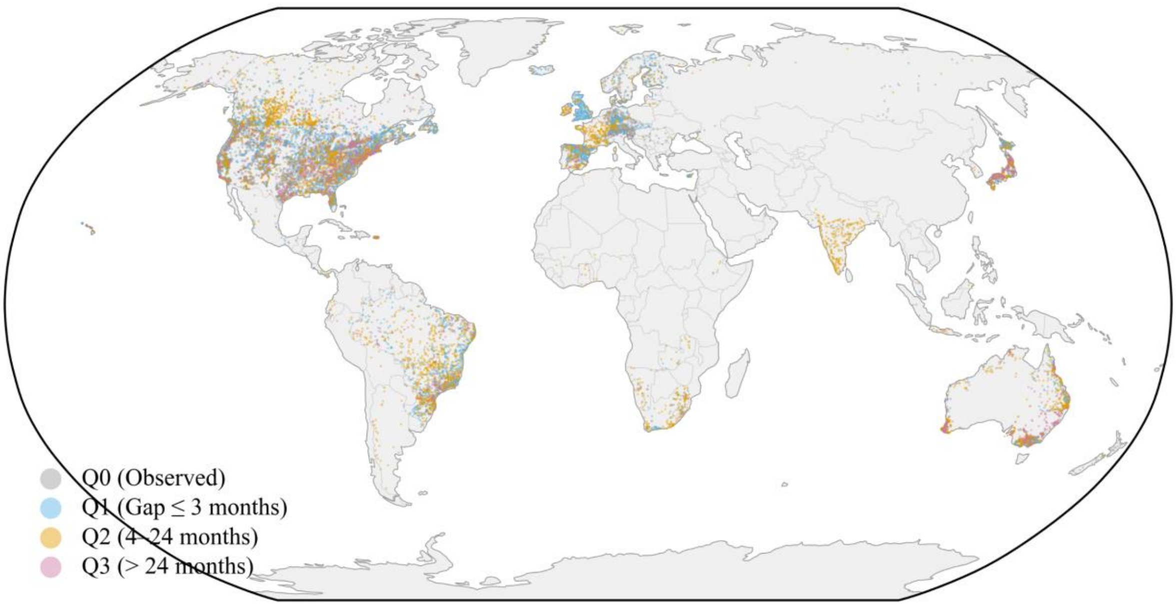


Figure 8. Global spatial distribution of GSIM-PLUS quality classes.

To verify whether GSIM-PLUS can maintain the hydrological continuity of data under complex conditions, we randomly selected five stations (Fig. 9) that illustrate characteristic product behaviors: AU_0002448 (temperate, short gaps with smooth connections to observations), BR_0001432 (tropical, seasonal recovery under medium mixed gaps), CA_0003714 (cold, restoration of winter gaps and snowmelt peaks), US_0009172 (complex mixed-gap structure), and AU_0002656 (arid, high variability and intermittent flow). No obvious discontinuities or unrealistic jumps appear between imputed and observed segments, indicating that GSIM-PLUS preserves seasonal signals and interannual variability with good hydrological realism.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

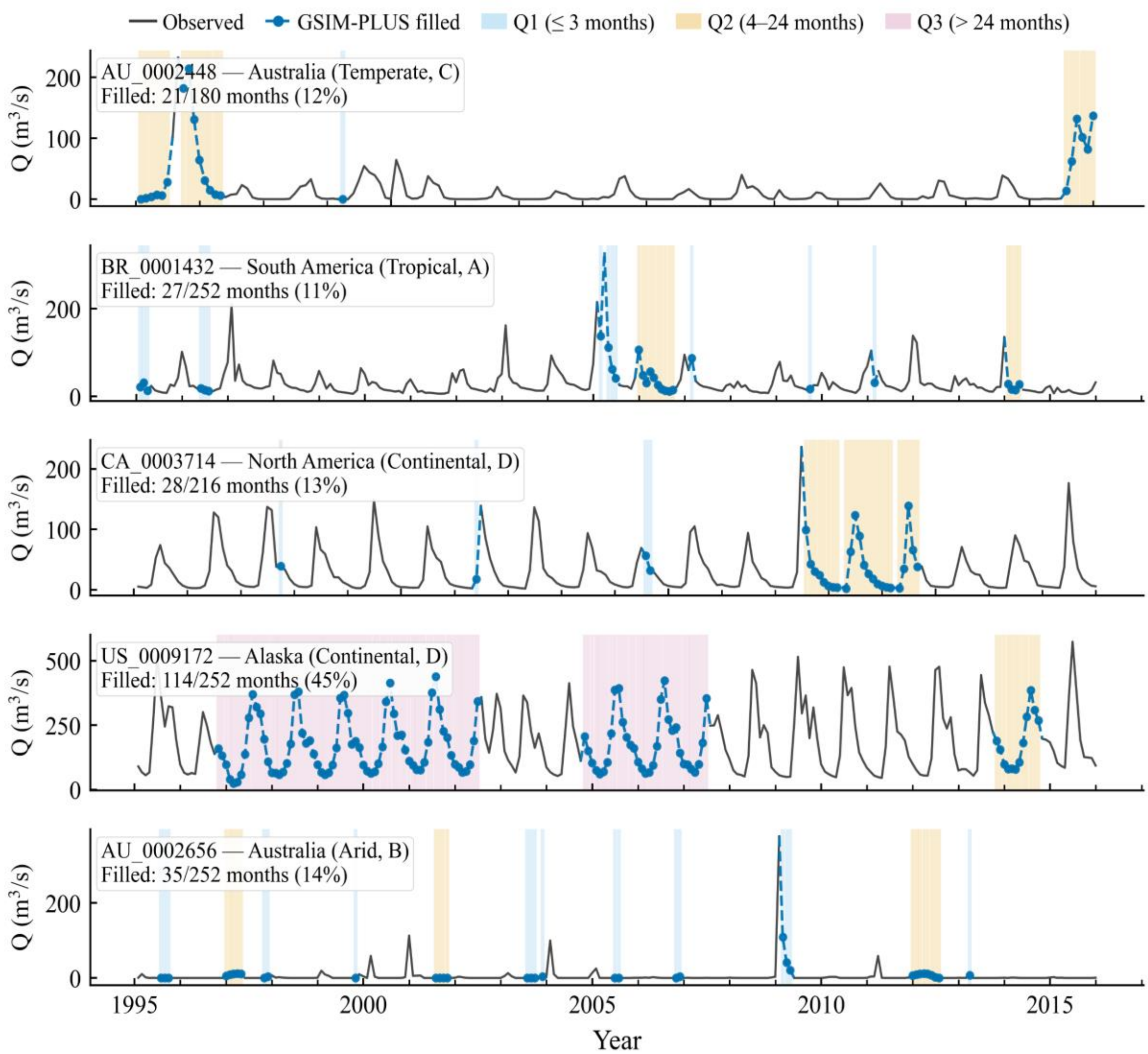


Figure 9. Representative station-level product examples.

5.3 Dataset composition and file organization

405 The public GSIM-PLUS release is distributed as a structured package that supports both direct hydrological use and full product traceability. The uploaded archive contains the main target-station product, the anchor companion product, the donor-similarity metadata, the parameter-sensitivity outputs, the core reproducibility scripts, and the release-level documentation files. The overall organization of the uploaded package is summarized in Table 5.

The release is designed around the practical use of the package as distributed, rather than around the internal workspace layout used during development. The main product entry point is "01_main_target_product_dtrr_guarded/", while "02_anchor_companion_product/" provides the companion archive for anchor stations. Supporting folders provide the donor-selection metadata, sensitivity-analysis outputs, and the main scripts required to audit or reproduce the published workflow. This package structure therefore makes GSIM-PLUS available not only as a filled monthly streamflow archive, but also as a traceable release in which users can recover both reconstruction provenance and the basin-context metadata associated with each record.

Table 5. Organisation of the GSIM-PLUS release package and the main function of each file group.

Component	Representative path	Main contents	Quantity or scope	Primary purpose
Main target-station product	GSIM_PLUS_release_package/ 01_main_target_product_dtrr_guarded/	Full monthly series, infilled-only table, station-level filling summary, product summary JSON, and station-split target files	8,731 target stations; 277,729 infilled monthly records; 1995-2015	Canonical GSIM-PLUS release for station-scale analysis, regional synthesis, and model evaluation
Station-split target series	GSIM_PLUS_release_package/ 01_main_target_product_dtrr_guarded/ GSIM_fill/	One CSV per target station with station_id, date, year, month, observed_streamflow, final_streamflow, segment_length, fill_method, quality_flag, kg_major, kg_code, arid_flag, low_flow_flag, context_risk_flag, and guard_applied	8,731 station files	Inspection of individual records, station-level diagnostics, and quality/context-based filtering
Anchor companion product	GSIM_PLUS_release_package/ 02_anchor_companion_product/	Anchor full-series table, infilled-only table, station-level filling summary, product summary JSON, and station-split anchor files	7,323 anchor stations; 25,542 infilled monthly records; 1995-2015	Companion archive for anchor consistency checks, reference-station analysis, and donor-network interpretation
Similarity metadata	GSIM_PLUS_release_package/ 03_similarity_metadata/	Target-to-anchor and anchor-to-anchor feature weights, Top-5 donor tables, and similarity summary files	Global matching metadata for both target and anchor workflows	Interpretation of donor selection and audit of the similarity-matching process
Parameter-sensitivity outputs	GSIM_PLUS_release_package/ 04_parameter_sensitivity_analysis/	Weight-sensitivity result directory, summary tables, key-result tables, interpretation files, and run_weight_sensitivity_analysis.py	Multiple weighting configurations and comparison outputs	Documentation of robustness to donor-weight perturbation
Reproducibility scripts	GSIM_PLUS_release_package/ 05_reproducibility_scripts/	Core scripts for donor matching, validation dispatch, target-product building, and anchor-product building	Key scripts spanning the released processing chain	Reproduction and audit of the published GSIM-PLUS workflow

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Component	Representative path	Main contents	Quantity or scope	Primary purpose
Release documentation	GSIM_PLUS_release_package/ README_release.md and PACKAGE_MANIFEST.md	Package overview, statistics, naming notes, source-path mapping, and release manifest	Release-level documentation	User guidance, package navigation, and mapping between release contents and workspace outputs

5.4 Dataset limitations and recommended use

Although GSIM-PLUS substantially improves the continuity of global monthly streamflow records, its applicability still has clear boundaries. First, uncertainty is higher for long continuous gaps than for short gaps. In particular, Q3 records are more suitable for regional comparison, long-term change detection, and exploratory analysis than for being treated as equivalent to original observations. Second, DTRR reconstruction at low-flow and intermittent-flow stations is more susceptible to instability under long-recursive settings, which is why the guarded fallback rule is applied to stations with median_flow < 0.02 m^3 s^{-1} during product generation.

For high-accuracy studies, we recommend prioritizing Q0 + Q1 records. For long-term trend analysis and regional comparison, Q2 can also be included. If all quality levels are used, the quality-flag composition should be reported and sensitivity analyses should be conducted jointly with segment_length and completeness changes. Overall, the key strength of GSIM-PLUS is that it systematically improves the continuity of global monthly streamflow data while preserving traceability to original observations and providing quality flags for tiered use.

It is worth noting that GSIM-PLUS has a complementary rather than substitutive relationship with gridded runoff products (e.g., GRUN) and reanalysis runoff outputs (e.g., ERA5-Land) in terms of application orientation. Gridded products provide spatially continuous runoff estimates, which are suitable for water resource assessments in ungauged regions. Reanalysis products deliver physically consistent long-term simulation series. In contrast, the core value of GSIM-PLUS lies in preserving the original characteristics of station observations — including the observed response characteristics at basin outlets and local hydrological signals — while improving temporal continuity through quality-controlled gap filling. For applications requiring station-scale in-situ data as a benchmark, such as hydrological model validation, trend detection, and cross-basin comparative analysis, GSIM-PLUS provides one of the most comprehensive global archives of monthly runoff stations currently available. In the future, the combined use of GSIM-PLUS station data

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

and gridded products is expected to further improve the accuracy and coverage of global hydrological assessments.

Here we would also like to clarify that a large number of stations were excluded during the station screening process. The 14,905 excluded stations with low data completeness were not uniformly distributed; instead, they were mainly concentrated in regions where the observational network has a short history or has suffered severe maintenance disruptions. This reflects the imbalance in global hydrological monitoring infrastructure, rather than coverage blind spots in the GSIM-PLUS framework. In these regions, the original GSIM archive itself lacks sufficient time series to support any statistical reconstruction method. In the future, as national observational networks are restored and satellite-retrieved runoff products become more mature, the coverage of GSIM-PLUS can be further expanded within the same methodological framework.

Data availability

The GSIM-PLUS data product is publicly available from Zenodo at https://doi.org/10.5281/zenodo.21425702 (Chen, 2026). The Zenodo archive contains the released gap-filled monthly streamflow dataset, including the target-station product, the anchor companion product, quality flags, contextual metadata, and associated summary files. The published data package corresponds to the GSIM-PLUS release described in this paper.

Code availability

The source code used to generate, validate, and document GSIM-PLUS is publicly available at https://github.com/skyhhu2024-create/GSIM-PLUS. The repository contains the scripts for station screening, feature construction, similarity-based donor selection, DTRR reconstruction, the MAML low-flow safeguard, validation experiments, product generation, and manuscript figure production. Large external input datasets are not redistributed through the repository and must be obtained from their original providers.

6 Conclusions

This study presents GSIM-PLUS, a gap-filled global monthly streamflow dataset for 1995–2015 derived from the GSIM monthly archive. By combining basin-similarity-based donor matching with station-level

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

recursive reconstruction, GSIM-PLUS fills 303,271 missing monthly records across 16,054 stations. This increases the median completeness of the target stations from 66.3 % to 81.0 % and that of the full station set from 86.9 % to 95.2 %, while preserving explicit information on reconstruction provenance and context.

The production workflow uses DTRR as the default reconstruction method, with a guarded MAML fallback for a limited subset of very-low-flow stations. Multi-scenario validation identified DTRR as the strongest overall method across scattered, continuous, mixed, and very long gaps. Performance remained strong for gaps of up to 12 months and retained positive, although reduced, skill for gaps exceeding 25 months. Comparison with independent GRDC observations at 16 co-located stations provided an external reference check, indicating that temporal co-variation was generally reproduced more reliably than exact flow magnitude, although agreement varied among stations.

The main contribution of GSIM-PLUS is therefore not simply the number of reconstructed values, but the provision of a more continuous global monthly streamflow archive in which reconstruction class, gap length, fill method, and major basin-context risks are explicitly documented. The product supports regional hydrological analysis, large-sample studies, model evaluation, and other monthly-scale applications requiring station records with improved temporal continuity. Reconstructed values should nevertheless be used more cautiously in arid basins, episodic or persistent very-low-flow conditions, and very long continuous gaps. Overall, GSIM-PLUS provides a practical and traceable foundation for broader reuse of global monthly streamflow observations.

Author contributions

MR C initiated this investigation. WH L, HC L, B Y, XZ X, JX Z, W D and J W designed the study. MR C developed the model code and performed the analysis. JY X and BY reviewed this manuscript and provided academic guidance.MR C prepared the paper with contributions from all co-authors.

Acknowledgements

We gratefully acknowledge the data providers and institutions that made this study possible, including the GSIM project, the Global Runoff Data Centre (GRDC), the ERA5-Land reanalysis team, and the HydroSHEDS database developers. Their efforts in data collection, curation, and open dissemination provided the essential foundation for the development of GSIM-PLUS. We also thank the editors and reviewers for their constructive comments and suggestions, which substantially improved the manuscript.

Financial support

This work was jointly supported by the National Key Research and Development Program of China (Grant No. 2024YFC3214801) and the sub-project of Xinjiang Major Science and Technology Special Project: Research on Key Technologies for Collaborative Regulation and Optimal Dispatching of multi-processes in irrigation canal water diversion, transmission and distribution (Grant No. 2022A02003-5-3).

Competing interests

The contact author has declared that none of the authors has any competing interests.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrol. Earth Syst. Sci.*, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Ali, G., Tetzlaff, D., Soulsby, C., McDonnell, J. J., and Capell, R.: A comparison of similarity indices for catchment classification using a cross-regional dataset, *Adv. Water Resour.*, 40, 11–22, https://doi.org/10.1016/j.advwatres.2012.01.008, 2012.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, *Hydrol. Earth Syst. Sci.*, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, *Water Resour. Res.*, 52, 3599–3622, https://doi.org/10.1002/2015WR018247, 2016.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I. J. M., and Wood, E. F.: Global fully distributed routing of streamflow to gauge locations, *Water Resour. Res.*, 56, e2019WR025341, https://doi.org/10.1029/2019WR025341, 2020.

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, *Water Resour. Res.*, 51, 4923–4947, https://doi.org/10.1002/2015WR017173, 2015.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H. (Eds.): Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, *Cambridge University Press*, Cambridge, UK, 2013.

Chen, M., Yong, B., Xu, X., Liao, W., Li, H., Zhang, J., Xiao, J., Dai, W., and Wang, J.: GSIM-PLUS: A Gap-Filled Global Monthly Streamflow Dataset for 1995–2015, Zenodo [data set], https://doi.org/10.5281/zenodo.21425702, 2026.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata, *Earth Syst. Sci. Data*, 10, 765–785, https://doi.org/10.5194/essd-10-765-2018, 2018.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive – Part 1: Station catalog and catchment boundary, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.887477, 2018.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, *Reviews of Geophysics*, 45, RG2004, https://doi.org/10.1029/2005RG000183, 2007.

Fekete, B. M., Robarts, R. D., Kumagai, M., Nachtnebel, H. P., Odada, E., and Zhulidov, A. V.: Time for in situ renaissance, *Science*, 349, 685–686, https://doi.org/10.1126/science.aac7358, 2015.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, *Earth Syst. Sci. Data*, 10, 787–804, https://doi.org/10.5194/essd-10-787-2018, 2018.

https://doi.org/10.5194/essd-2026-316
Preprint. Discussion started: 24 July 2026

Open Access Earth System Science Data Discussions

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

Gyau-Boakye, P. and Schultz, G. A.: Filling gaps in runoff time series in West Africa, *Hydrol. Sci. J.*, 39, 471–487, https://doi.org/10.1080/02626669409492769, 1994.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, *Hydrol. Process.*, 25, 1191–1200, https://doi.org/10.1002/hyp.7794, 2011.

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, *Earth Syst. Sci. Data*, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.

Harvey, R. J., Dixon, H., and Hannaford, J.: The spatial interpolation of daily streamflow records, *Water Resour. Res.*, 48, W12521, https://doi.org/10.1029/2012WR012270, 2012.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V.: Theory-guided data science: A new paradigm for scientific discovery from data, *IEEE Trans. Knowl. Data Eng.*, 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, *Hydrol. Earth Syst. Sci.*, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kundzewicz, Z. W., Graczyk, D., Maurer, T., Pińskwar, I., Radziejewski, M., Svensson, C., and Szwed, M.: Trend detection in river flow series: 1. Annual maximum flow, *Hydrol. Sci. J.*, 58, 797–810, https://doi.org/10.1080/02626667.2013.792205, 2013.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, *Hydrological Processes*, 27, 2171–2186, https://doi.org/10.1002/hyp.9740, 2013.

Merz, R., Blöschl, G., and Parajka, J.: Spatio-temporal variability of event runoff coefficients, *J. Hydrol.*, 312, 39–51, https://doi.org/10.1016/j.jhydrol.2005.02.014, 2005.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Stationarity is dead: Whither water management?, *Science*, 319, 573–574, https://doi.org/10.1126/science.1151915, 2008.

Mishra, A. K. and Coulibaly, P.: Developments in hydrometric network design: A review, *Rev. Geophys.*, 47, RG2001, https://doi.org/10.1029/2007RG000243, 2009.

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

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial interpolation of daily runoff: why physical similarity is better than spatial proximity, *Water Resour. Res.*, 44, W12423, https://doi.org/10.1029/2008WR006840, 2008.

Vörösmarty, C. J., Askew, A., Grabs, W., Barry, R. G., Birkett, C., Döll, P., Goodison, B., Hall, A., Jenne, R., Kitaev, L., Margules, J., Meier, M. F., Morel, R., Prowse, T., Schlosser, C. A., Serreze, M., Shiklomanov, I., and Yang, D.: Global water data: A newly endangered species, *Eos Trans. AGU*, 82, 54–58, https://doi.org/10.1029/01EO00031, 2001.

.